

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

<http://go.warwick.ac.uk/wrap/59616>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

# **ASPECTS OF FORECASTING AGGREGATE AND DISCRETE DATA**

**KLAUS LEITE PINTO VASCONCELLOS**

Thesis submitted for the degree of Doctor of Philosophy

**DEPARTMENT OF STATISTICS  
UNIVERSITY OF WARWICK  
DECEMBER 1992**

# CONTENTS

Summary	vi
CHAPTER 1. INTRODUCTION	
1.1 Nature of the problem	1
1.2 Aggregation and choice of the sampling interval	2
1.3 The problem of initialization	3
1.4 Scope of the thesis	3
CHAPTER 2. BASIC THEORY	
2.1 Introduction	6
2.2 The univariate DLM	
2.2.1 Definition	6
2.2.2 Updating and forecasting	7
2.2.3 Time Series DLM's	8
2.3 The first order constant model	
2.3.1 Definition and updating	9
2.3.2 Limiting behaviour	10
2.3.3 Discount factors	11
2.4 Polynomial trend models	
2.4.1 General polynomial model	12
2.4.2 Second order polynomial DLM	13
2.5 Models with unknown observational variance	17
2.6 Reference Analysis of the DLM	
2.6.1 Introduction	19
2.6.2 Updating equations	19
2.7 Dynamic Generalised Linear Models	
2.7.1 Motivation	22
2.7.2 Exponential family distribution	22
2.7.3 Definition of DGLM	23
2.7.4 Updating of DGLM	24
2.8 Related work	29
CHAPTER 3. REFERENCE ANALYSIS OF THE DLM	
CHAPTER 4. THE BOOKING MODEL	
4.1 Introduction	41
4.2 The booking problem	41

4.3 Model definition	44
4.4 Simple updating	45
4.5 The choice of $f(\cdot)$	48
4.6 Robustness of $f(\cdot)$	50
4.7 The updating problem	50
4.8 The general updating	54
4.9 Simulation Results	56
4.10 Conclusions	64
4.11 Appendix	
PART 1. Basic properties of both distributions	65
PART 2. Equating mean and variance	66
PART 3. $L_2$ minimization	67
PART 4. Comparison of the two methods and conclusions	70
 CHAPTER 5. DATA AGGREGATION	
5.1 Introduction	75
5.2 The constant first order polynomial DLM $\{1, 1, V, W\}$ case	
5.2.1 The $\{1, 1, V^*, W^*\}$ representation	77
5.2.2 The sophisticated model	79
5.3 The $\{F, I, V, W\}$ case	84
5.4 The $\{1, \lambda, V, W\}$ case	
5.4.1 Basic ideas	87
5.4.2 The $\{1, \lambda^n, V^*, W^*\}$ representation	88
5.4.3 Model sophistication	96
5.4.4 The importance of correlation	103
5.4.5 Misspecification of $\lambda$	113
5.5 General problem and linear growth	
5.5.1 General idea	115
5.5.2 The linear growth model	117
5.6 Aggregation and the booking model	123
5.7 Conclusions	124
 CHAPTER 6. CONCLUSIONS	
6.1 Review of Chapter 3	127
6.2 Review of Chapter 4	127
6.3 Analysis of results	129
6.4 Review of Chapter 5	131



## ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Prof. Jeff Harrison. His constant attention and wise advice were fundamental for this thesis to be completed, while my talks with him always gave me strong motivation to work. Also, my talks with Jeff were very important to improve my understanding of statistics, particularly of forecasting.

This PhD course was sponsored by the Brazilian governmental institution Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), whom I would like to thank for their financial support.

I am extremely indebted to my family in Brazil. Their patience and the support they gave me proved to be essential for this thesis to be finished. Also, I would like to thank my friends whose support towards the completion of this work was essential.

## SUMMARY

This work studies three related topics arising from the problem of forecasting airline passenger bookings.

The first topic concerns the initialization through the starting prior for a DLM (Dynamic Linear Model) or Generalized DLM. An approach is given which uses the first observations of the series much more efficiently than that suggested by Pole and West. Proper marginal priors are derived for stationary model components and proper marginal priors may be obtained for parameter subspaces and used for forecasting within that subspace well before a full proper prior is available.

The second topic proposes a model to forecast the number of people booking tickets for particular flights. The model is more realistic than those which are classically used, since it is a dynamic model and acknowledges discrete distributions. The basic idea is given by the Dynamic Generalized Linear Model and a key feature is given by the gamma to log-normal approximation that is developed.

The third topic consists of a study of temporal aggregation of a process that can be represented by a DLM. We give representation results for the simplest univariate cases, reveal some surprising phenomena, such as drastic model simplification with aggregation, and discuss some advantages and disadvantages of using the aggregated observations, depending on the forecasting objectives, as well as the importance of aggregation in our particular booking problem.

# CHAPTER 1

## INTRODUCTION

### 1.1 *Nature of the problem*

The basic motivation for the work presented here lies in the necessity of constructing a model to forecast airline passenger bookings. This constitutes an important practical application since the airline companies have their policy of passenger services almost entirely based on advance reservations. Therefore, it is very important for the company to have a precise idea of how the reservation process takes place. This will allow the company to take better decisions, as, e.g., optimal allocation of planes in terminals, possible increase of the load factor for some particular flights, reduction in the cost of oversales, a more efficient timetable, optimal allocation of seats in each plane, among other benefits, therefore, increasing profit and reducing costs.

The problem of producing a good estimate for future flight reservations has been studied to some extent by most airline companies (see Rothstein, 1985). However, the methodology which has been used by some companies for forecasting purposes may be subject to criticism in the choice of model and also in the use of data. Data used in the analysis is always the number of passengers who effectively take seats in the plane, rather than the actual number of booked tickets for the respective flight. Recording this last data should give airline companies information concerning the real demand for the flights. Data has been used, therefore, as an input to a controlled system, when the real input should be the number of tickets which were initially booked, or better still requested.

It is important to notice that the number of booked tickets is data *subject to control*. If the company, for example, is forced to close a particular class, then, this will be reflected in the booking of tickets, and, consequently, will influence the total number of people taking the flight. This important fact has been neglected by people who propose forecasting models for flights. The available models are, therefore, fed with truncated data, and relevant information is not taken into account.

Correction of data is a fundamental factor for a good performance of models and it is important to note all special events that may influence the booking of tickets. Special events may affect the usual behaviour of data and it will be probably a good idea to consider intervention, since we are dealing with the case where data deviates from its regular pattern.

It is, then, very clear the importance of collecting information about the demand for flights; the company loses by using only the net bookings and disregarding can-



cellation data. This data would also provide information about the control to which the input is being submitted, i.e., decisions that are being taken about redistribution of places in the planes, closing of specific terminals, overbookings, etc. It is suggested, therefore, that the reservation data must be collected along with cancellations of booked places; the model to be built to explain the number of passengers in the plane must take that last factor into account. Here, it is also essential for the forecaster to have the facility of informing the model about the control described above as well as special events; the fraction of class open and decisions on closing specific terminals is worth to be noting. The model should include a prior distribution, which must be derived from information given by experts, etc. We will assume for the development of the model that we are already given a prior distribution.

Probability assumptions about data must be considered very carefully. Some airline companies erroneously use normal distributions, or truncated normal distributions (Lee, 1988), a questionable approach, since we are dealing with small numbers, and, therefore, great skewness. This is particularly true if we consider club class, where fewer people will book tickets. Also, we will be dealing with small numbers (and therefore great skewness) if the bookings are made very near to date of departure of the respective flight. It is, then, much more realistic to work with *discrete* distributions and acknowledge great skewness of data. For this reason, the model we develop here fits a discrete forecasting distribution for the future demand of seats.

### 1.2 *Aggregation and choice of the sampling interval*

A very important component in the model is certainly given by the length of the sampling interval, where we define an observation for a given period as the total number of ‘counts’ corresponding to that time interval. In other words, data is cumulative over the periods. It is important to consider the fact that in any application of a forecasting procedure, the sampling period, as defined above, plays a basic role, constituting a fundamental factor determining what are the relevant effects influencing the behaviour of the process in study. If data is sampled over a very long period, i.e, if the aggregation level is too high, then, a large amount of information may be aggregated, and we will probably lose a detail level that would be necessary for good decision making. But, on the other hand, if data is sampled too frequently, the high frequency components will dominate, and present a serious modelling problem for the practitioner in detecting those important systematic components that may be needed to produce good forecasts. It is clear, then, that the sampling period must be chosen consistently with the forecasting objectives, and this must be perfectly

defined before we decide what forecasting method must be employed.

In our specific problem, the sampling period is certainly a crucial element, since data regarding the passenger demand must be seasonal and subject to much variation, and the optimal sampling period may well vary, depending on the specific flight we are studying. For example, for some flights, it may be a reasonable approach to consider the number of reservations, for a specific flight departing on a specific date, as independent random variables. For other flights, there may be a correlation between different weeks, but the hypothesis of independence may be reasonable for a larger sampling interval. It is, therefore, very important to define precisely what our main forecasting objectives are, and choose the sampling interval accordingly.

### 1.3 *The problem of initialization*

The model we construct to deal with the booking data has its base in the Dynamic Generalized Linear Model (West and Harrison, 1986; West, Harrison and Migon, 1985). One of the main features of this model is that, for each instant  $t$  of time, we consider a state vector parameter  $\theta_t$ , such that the sequence  $\{\theta_t\}$  obeys a Markovian evolution. Then, we consider a linear function  $\lambda_t = F_t' \theta_t$  of this state parameter (where  $F_t$  is a known vector or regressors), which is linked with the mean of the process, through a known bijection (in our specific case, this bijection will be an exponential function). The distribution for  $\theta_t$  is updated sequentially, as we collect information along time. This methodology requires, therefore, a prior distribution for the state parameter, and there may be occasions when we will depart from a rather vague, or uninformative, prior and construct a proper prior using the information provided by the first observations of the process. Thus, it will be important to consider a method that will efficiently use these first observations in order to obtain a useful initial prior distribution for the state space vector.

### 1.4 *Scope of the thesis*

The organization of this work is as follows.

In Chapter 2 we review the theory of the Dynamic Linear Models, presenting the main ideas that will be necessary for the development in the next three chapters.

In Chapter 3, we discuss the reference analysis of the DLM. This relates to obtaining a proper distribution for the state parameter, which is necessary for the initialization of the forecast system. The approach which is proposed by Pole and West (1989) and is given in West and Harrison (1989) presents some drawbacks. The first one is that it does not use information that can be supplied by the model in order to find an initial proper distribution that can be available, at least for

a particular subspace of the state space. It also supposes the system matrix to be non-singular, hence, we cannot apply the method more generally, for example, in the case of simple moving average processes. A third drawback is that it does not take into account the fact that if we are interested in obtaining forecasts for a particular horizon, where we will always be in a certain subspace, then, a proper distribution for the full space is not needed. As soon as we can obtain a conditional proper forecast distribution for the subspace in question, this conditional distribution will be sufficient for our purposes. In Chapter 3 we present an alternative approach for the initialization of the DLM.

In Chapter 4, we construct our model, which is based on the Dynamic Generalized Linear Model, to forecast the number of booked tickets for particular flights taking off at regular intervals of time. This model is more realistic than the model that has been used for the same purpose, acknowledging great skewness of the distributions. A very important feature of the developed model is given by the gamma to lognormal approximation that is developed. This density approximation seems to work very well in the specific problem we consider.

In Chapter 5, we discuss the problem of aggregation. As was mentioned before, an important decision related to the use of data consists in the sampling interval that must be chosen. If data is sampled too frequently, high frequency effects will become dominant and this can lead us to unnecessary complications when we consider an explanatory model. It is important to observe that sometimes we cannot find a simple model for a very small sampling interval, but one will emerge for larger intervals. For example, if data is collected daily, then, the specific day of the week for which each observation is collected can constitute an important explanatory factor. But, if we consider the number of booking tickets for the overall week, this factor will disappear. This can provide a huge model simplification, since we are removing up to six parameters. These ideas lead us to the conclusion that it may be worth working with aggregated data, where we consider the total number of observations for the larger sampling period, rather than the observations for the smaller periods, individually.

The problem of aggregation of observations as well as the dependence of the model with the sampling interval has been extensively studied in the literature. Amemiya and Wu (1972) investigated the purely autoregressive models and Brewer (1973) examined the effect of aggregation for ARMA models. Both studies only treat stationary processes or derivable stationary ARIMA processes. Tiao (1972) considers aggregation for the  $IMA(d, q)$  model and Wei (1978) extends the results

for the general multiplicative seasonal model  $ARIMA(p, d, q) \times (P, D, Q)_s$ , giving results concerning the  $ARIMA$  representation for the aggregated observations. For a purely autoregressive process, Stram and Wei (1986) give the exact order for the aggregate model and also show that aggregation can reduce the autoregressive order of  $AR(p)$  and  $ARIMA(p, d, q)$ . They also observe that the  $IMA(d, q)$  model can be reduced to a simple  $IMA(d, 0)$  by aggregation. Also, Gonzales (1992) examines the gains in accuracy when a series is sampled at more frequent intervals.

In Chapter 5, we begin our investigation by studying the problem of aggregation for the DLM. We consider representation results for the more simple models. For those models we obtain a set conditions, which, if satisfied, allows us to represent the aggregated data by a model which is, at least, as simple as the original one. We also show that, in some particular cases, we can obtain a simpler model by data aggregation. The model reduction can be even more drastic than those indicated by Stram and Wei (1986). For example, we can reduce an  $ARMA(1, 1)$  process to an  $AR(1)$ , by aggregating, or an  $IMA(2, 2)$  model (linear growth model) can be reduced to the very simple  $IMA(1, 0)$  structure. Such results work as an indication that there must always exist an optimal aggregation level, an optimal sampling interval that we must use, depending of what our forecasting purposes are. We show that the inclusion of a zero eigenvalue can always provide us with a representation for the aggregated model and, for the more simple cases of the constant DLM, the theoretical and practical consequences of this inclusion are discussed. The extension of the model constructed in Chapter 4 to work with aggregated data is also considered in this chapter.

Finally, in Chapter 6 we present the main conclusions of these chapters, and make suggestions for further study.

## CHAPTER 2

### BASIC THEORY

#### 2.1 Introduction

In this chapter we review the basic theory related to the Dynamic Linear Models (Harrison and Stevens, 1971, 1976; West and Harrison, 1989) that will be necessary for the discussion of the ideas in the next chapters. The general univariate DLM is briefly discussed, and we give particular attention to the constant models. We analyse briefly the simplest univariate cases, which correspond to the first and second order polynomial models. The case of the unknown observational variance is considered and we briefly discuss the variance learning procedure in the simplest cases. Then, we discuss the initialization of the forecast system, presenting the method proposed by Pole and West (1989) to obtain a proper distribution for the state vector based on the first observations of the process. Also, we introduce the basic ideas related to the generalized dynamic linear model. At the end of the chapter, we give some references related with what is presented in this brief review. The main reference for the chapter is West and Harrison (1989), where a detailed discussion of the topics presented here can be found.

#### 2.2 The univariate DLM

##### 2.2.1 Definition

Let  $y_t$  represent the process for which we collect information at each instant  $t$  of time. We relate  $y_t$  to a quantity  $\theta_t$  via a dynamic linear regression of the form

$$y_t = F_t' \theta_t + v_t \quad (2.1)$$

where  $F_t$  is a regression vector of independent variables, which is known for each instant  $t$  of time and  $\{v_t\}$  is a random sequence of independent errors uncorrelated with  $\theta_t$ . We suppose  $\{v_t\}$  is a zero mean normally distributed sequence, such that  $v_t \sim N[0, V_t]$ , for each  $t$  and the sequence  $\{V_t\}$  of variances is also supposed to be completely known. The quantity  $\theta_t$  here plays the role of a dynamic vector of regression parameters, which is frequently referred to as the *state vector* of the model. The state vector contains, therefore, the relevant parameters we need at time  $t$  to express our beliefs about  $y_t$  in the sense that  $(y_t | \theta_t) \sim N[F_t' \theta_t, V_t]$  is independent of past values of the process. The state vector  $\theta_t$  has a one-step Markov evolution, its time behaviour being governed by the equation

$$\theta_t = G_t \theta_{t-1} + \omega_t \quad (2.2)$$

where  $G_t$  is the *system matrix* of the model, assumed known for each instant  $t$  of time and  $\{\omega_t\}$  is also a random sequence of independent errors such that  $\omega_t$  is uncorrelated with  $\theta_{t-1}$ . We suppose, as with  $\{v_t\}$ , that  $\{\omega_t\}$  is a zero mean normally distributed sequence, such that  $\omega_t \sim N[0, W_t]$ , for each  $t$ , the sequence  $\{W_t\}$  of covariance matrices being completely known. We call (2.1) the *observation equation* of the model, and  $v_t$  is the observational error. Also, (2.2) will be called the *system equation* of the model, and  $\omega_t$  is the evolution error, assumed independent of  $v_t$ . The univariate Dynamic Linear Model can, then, be characterised by a quadruple

$$\{F, G, V, W\}_t = \{F_t, G_t, V_t, W_t\}$$

which is known for each instant  $t$  of time. Let  $D_t$  represent all the relevant information we have up to the time  $t$ . Then, we consider the above quadruple, together with the initial information  $(\theta_0|D_0) \sim N[m_0, C_0]$  for some moments  $m_0$  and  $C_0$ . We also assume this initial distribution for the state to be independent of  $v_t$  and  $\omega_t$ . Then, we have the univariate DLM completely defined.

### 2.2.2 Updating and forecasting

If we assume that for each instant  $t$  of time, we will have  $D_t = \{y_t, D_{t-1}\}$ , which means, we never have external information available, then, with  $\{F, G, V, W\}_t$  known for all  $t$ , the DLM can be updated as follows:

- 1) At time  $t - 1$  we have a posterior distribution for the state vector, given as

$$(\theta_{t-1}|D_{t-1}) \sim N[m_{t-1}, C_{t-1}]$$

for some mean  $m_{t-1}$  and covariance matrix  $C_{t-1}$ .

- 2) The prior distribution for the state vector at time  $t$  is, from (2.2):

$$(\theta_t|D_{t-1}) \sim N[a_t, R_t]$$

where  $a_t = G_t m_{t-1}$  and  $R_t = G_t C_{t-1} G_t' + W_t$ .

- 3) Using (2.1), we obtain the one-step ahead forecast distribution

$$(y_t|D_{t-1}) \sim N[f_t, Q_t]$$

where  $f_t = F_t' a_t$  and  $Q_t = F_t' R_t F_t + V_t$ .

- 4) We now consider the joint normal bivariate distribution of  $y_t$  and  $\theta_t$ , from which we can derive the posterior distribution at time  $t$  for the state vector, given by:

$$(\theta_t|D_t) \sim N[m_t, C_t]$$

with  $m_t = a_t + A_t(y_t - f_t)$  and  $C_t = R_t - A_t Q_t A_t'$ , where  $A_t = R_t F_t Q_t^{-1}$ .

The procedure above puts the updating in closed form, and we can, at any instant  $t$ , produce forecasts for the next observations of the process under study. In fact, we very simply obtain the  $k$  step ahead distributions for the state vector and the original process, as

$$(\theta_{t+k}|D_t) \sim N[a_t(k), R_t(k)]$$

$$(y_{t+k}|D_t) \sim N[f_t(k), Q_t(k)]$$

where the moments of the distributions can be obtained as

$$f_t(k) = F_t' G_{t+k} G_{t+k-1} \dots G_{t+1} m_t$$

$$Q_t(k) = F_t' R_t(k) F_t + V_{t+k}$$

where

$$R_t(k) = G_{t+k} R_t(k-1) G_{t+k}' + W_{t+k}$$

with initial value  $R_t(0) = C_t$ .

### 2.2.3 Time Series DLM's

A very important subclass of DLMs are the *Time Series* DLMs, or TSDLMs, those which are characterized by *constant* (not changing with time)  $F$  and  $G$ . A TSDLM where the variances  $V$  and  $W$  are also constant with time is referred to as a *constant* DLM. In fact, most of the classic linear time series models that have been extensively used in the literature can be put in this framework, which makes the constant DLM a subset of particular interest in the general class.

A fundamental concept in the study of the DLM is the one of observability. Let  $n$  be the dimension of the state vector. The *observability matrix*  $T$ , for the TSDLM, is, by definition, the matrix

$$T = \begin{pmatrix} F' \\ F'G \\ \vdots \\ F'G^{n-1} \end{pmatrix}$$

We consider a TSDLM to be *observable* if and only if the above matrix is non-singular. Briefly explaining, an observable TSDLM means that the observations we collect along time provide information about all the components of the state vector. Observe, for example, that if we have a deterministic evolution equation (the variance of the evolution error being identically zero), then, the path of the state vector can be completely determined from the  $n$  first values of the mean response

$\mu_t(k) = F'G^k\theta_t$ , using the observability matrix. In fact, the rank of this matrix gives us the dimension for which we have an observable model with the same  $\mu_t(k)$ . It is important to observe that if the system is observable, then, the forecast function  $f_t(k) = F'G^k m_t$  can be completely determined by its first  $n$  values.

A fundamental result states that if we have an observable constant DLM with finite observational and evolution variances, then the sequence  $\{C_t\}$  of the posterior covariance matrices of the state will be convergent to a finite matrix (Harrison (1985), West and Harrison (1989)). This allows us to know, a priori, before any observation is available, the limiting behaviour of the forecasting expressions, simply by calculating the limit value of the quantities that appear in the updating equations.

## 2.3 The first order constant model

### 2.3.1 Definition and updating

The most simple form of the DLM is the first order constant model, the observation and evolution equations being respectively defined by

$$\begin{cases} y_t = \theta_t + v_t & (2.3) \\ \theta_t = \theta_{t-1} + \omega_t, & (2.4) \end{cases}$$

where  $Var[v_t] = V$  and  $Var[\omega_t] = W$ , with known  $V$  and  $W$ .

Then, it is easily seen that the updating steps 1 to 4 of the anterior section are trivially reduced as follows:

- 1) At time  $t - 1$  the posterior distribution for the state is

$$(\theta_{t-1}|D_{t-1}) \sim N[m_{t-1}, C_{t-1}]$$

for some mean  $m_{t-1}$  and variance  $C_{t-1}$ .

- 2) The prior distribution for the state at time  $t$  is:

$$(\theta_t|D_{t-1}) \sim N[a_t, R_t]$$

where  $a_t = m_{t-1}$  and  $R_t = C_{t-1} + W$ .

- 3) One-step ahead forecast distribution:

$$(y_t|D_{t-1}) \sim N[f_t, Q_t]$$

where  $f_t = a_t$  and  $Q_t = R_t + V$ .

- 4) Posterior distribution at time  $t$  for the state:

$$(\theta_t|D_t) \sim N[m_t, C_t]$$

with  $m_t = a_t + A_t(y_t - f_t)$  and  $C_t = A_t V$ , where  $A_t = R_t/Q_t$ .



### 2.3.2 Limiting behaviour

Although practically useful, the constant model is somewhat restricted, since it is not open to any external information the modeller would like to transmit. However, the simplicity of the model allows us to derive important results that can be used for developing ideas concerning the more sophisticated models. A crucial role is played by the variance ratio  $r = V/W$  which measures variation of the observations with respect to the systematic variation (this quantity is typically around 20 for most part of applications). Let  $A, C, Q, R$  denote the limit values of  $A_t, C_t, Q_t, R_t$ , respectively. Then, from step 2 above we get  $R = C + W$ . But, from step 4, we get  $R = AQ$ , and, from steps 2 and 3 we get  $R = A(C + V + W)$ . Comparing these two expressions for  $R$ , we get

$$C + W = A(C + V + W)$$

and, since, from step 4,  $C = AV$ , we get

$$AV + W = A(AV + V + W)$$

which can be divided by  $W$ , being rewritten as

$$Ar + 1 = A(Ar + r + 1)$$

giving

$$rA^2 + A - 1 = 0$$

and, since the roots have different signs, we must take the biggest one, arriving at

$$A = \frac{2}{1 + \sqrt{1 + 4r}} \quad (2.5)$$

and, from above, we readily obtain  $C, Q, R$ .

We must observe that, in the limit, the updating equation for  $m_t$  in step 4, will be written in the form

$$m_t = m_{t-1} + A(y_t - m_{t-1}) = Ay_t + \delta m_{t-1} \quad (2.6)$$

where  $\delta = 1 - A$ . More formally, it can be seen that  $m_t - m_{t-1} - Ae_t$  will converge in probability to zero. So, in the limit, the presented approach will be equivalent to the point predictor of Holt (1957), the exponentially weighted regression of Brown (1962) and is contained in the ARIMA(0,1,1) representation of Box and Jenkins (1976). A fundamental point in all of these methods is that we are obtaining the

one-step ahead forecast for the next period as a weighted average of the forecast we had for  $y_t$  and the actual value of  $y_t$ . Since  $A$  represents the weight given to the observation, we must expect that  $A$  will be large (that is, almost one) if we have a very small uncertainty about the observation compared to the systematic variation of the mean level  $\theta_t$ . If, on the other hand,  $y_t$  is subject to a lot of variation, then, we will give more importance to the last estimator we had,  $m_{t-1}$ . This is confirmed by (2.5) since, from this expression, we can readily see that  $A$  will be near one for very small values of  $r$ , and will be near zero, if  $r$  is quite large. We can also think that, when  $V = 0$ , we will have  $y_t = y_{t-1} + \omega_t$ , which means that the best estimator for the next observation is given by the last data. This is obtained for  $A = 1$ . On the contrary, when  $W = 0$ , we have  $y_t = \theta + v_t$ , where  $\theta$  is a deterministic constant level. The best estimator will always be, then, the value of the level  $\theta$ . Therefore, we should have, in this case,  $A = 0$ . From (2.6) it must be clear that the value of  $A$  determines the sensitivity of the predictor to the most recent observations of the process. The larger the value of  $A$  the larger the sensitivity becomes.

### 2.3.3 Discount factors

We observe that, in the limit

$$R = C + W = C + \frac{V}{r} = \left(1 + \frac{1}{Ar}\right) C = \frac{C}{\delta}$$

and  $W = AC/\delta$  is a fixed proportion of  $C$ . Therefore, in the limit, the increase in the variance of the state parameter due to the error  $\omega_t$  is equivalent to an increase of this same variance of a fixed proportion  $\delta^{-1}$ . Since convergence is usually very fast for the constant model, we can, in practice, consider the option of adopting a constant *discount factor*  $\delta$ , instead of specifying a constant variance  $W$  in the evolution equation. That means, in the second step of the updating, the prior distribution for time  $t$  will be obtained as  $R_t = C_{t-1}/\delta$ , instead of adding a constant variance  $W$ . Observe that the limiting behaviour of this new model is the same as the constant one. This can be directly verified from the expression for  $C_t^{-1}$ , which, from step 4, is

$$C_t^{-1} = V^{-1} R_t^{-1} Q_t = V^{-1} R_t^{-1} (R_t + V) = V^{-1} + R_t^{-1} = V^{-1} + \delta C_{t-1}^{-1}$$

and, therefore

$$C_t^{-1} = V^{-1}(1 + \delta + \dots + \delta^{t-1}) + \delta^t C_0^{-1}$$

which gives, in the limit,  $C^{-1} = V^{-1}/(1 - \delta)$ , or  $C = AV$ , as we had before. These are the basic ideas of discounting, which have been extensively studied, the approach

being introduced in Ameen and Harrison (1985) and described in practical detail in Harrison and West (1987) and Harrison (1988). This discount factor idea gives us a very useful alternative approach, since the variance  $W$  in the evolution equation is, in practice, difficult to specify.

## 2.4 Polynomial trend models

### 2.4.1 General polynomial model

We extend the ideas of the last section in order to consider a general class of time series models that have been extensively used in practice, through many years. The polynomial trend models basically use the idea that a well behaved function can be locally approximated by a low order polynomial with a good precision. The most simple is, of course, the constant trend model, which is the case we have just discussed. The next simplest case is the linear growth model which has been widely studied in the literature (Harrison, 1965, 1967; Godolphin and Harrison, 1975). We present the general construction, although, in practice, it is very rare that a third or higher degree polynomial will be used.

We define the  $n^{th}$  order polynomial models as the particular class of observable TSDLM for which the forecast function  $f_t(k) = F'G^k m_t$  can be written as

$$f_t(k) = a_{t0} + a_{t1}k + a_{t2}k^2 + \cdots + a_{t,n-1}k^{n-1} \quad (2.7)$$

for all  $t, k \geq 0$ .

Such polynomial forecast functions are discussed in Harrison (1965, 1967), theoretical aspects being explored in Godolphin and Harrison (1975). In Godolphin and Stone (1980) this polynomial form of the forecast function is generalised to allow the first few values to be irregular.

In fact, it can be seen that the form of any given forecast function  $f_t(k)$ , regarded as a function of  $k$ , is determined by the eigenvalue structure of the system matrix  $G$ . In other words, two observable TSDLM's will have the same form of forecast function if and only if their system matrices have the same set of eigenvalues. In the particular case of (2.7) above, the system matrix must be similar to the  $n^{th}$  order Jordan block with unit eigenvalue  $J_n(1)$ . We denote

$$J_n(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}.$$

and, in the case of (2.7), there must exist a non-singular matrix  $H$  such that  $HGH^{-1} = J_n(1)$ . The generalisation in Godolphin and Stone (1980) corresponds to extend the system matrix with the inclusion above of a  $J_k(0)$  block.

It is common, in practice, to work with the *canonical*  $n^{th}$  order polynomial DLM, defined as that one for which  $G = J_n(1)$  and  $F = E_n = (1, 0, \dots, 0)'$ , the first element in the canonical ordered set of coordinate vectors of the  $R^n$ . In this model, the first coordinate of the state vector gives the level of the process at time  $t$ , and, for  $1 < j \leq n$ , the  $j^{th}$  coordinate of the state vector represents the systematic change in the  $(j - 1)^{th}$  coordinate.

#### 2.4.2 Second order polynomial DLM

The most simple observable model that constitutes an extension from the first order polynomial DLM is, naturally, the canonical second order polynomial DLM. For this model, we will have a two dimensional state vector with system matrix given by  $G = J_2(1)$  and constant regression vector  $F = E_2 = (1, 0)'$ . We define  $\theta'_t = (\mu_t, \beta_t)$ , and write the model equations as

$$y_t = \mu_t + v_t \quad (2.8)$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \omega_{t1} \quad (2.9)$$

$$\beta_t = \beta_{t-1} + \omega_{t2} \quad (2.10)$$

where the errors above are zero mean normally distributed,  $\omega_{t1}$  and  $\omega_{t2}$  are uncorrelated with  $v_t$  and we have  $Var[v_t] = V_t$ ,  $Var[\omega_{t1}] = W_{t1}$ ,  $Var[\omega_{t2}] = W_{t2}$  and  $Cov[\omega_{t1}, \omega_{t2}] = W_{t3}$ .

We can apply the updating procedure of the last section in order to obtain the forecast distributions for the process at each instant  $t$ . It is important to discuss here the limiting behaviour of the constant model. For simplicity, we consider that the variances are scaled by the observational variance. If  $V_t = V = 1$ , and  $W_{t1}$ ,  $W_{t2}$  and  $W_{t3}$  are constant variances given by  $W_1$ ,  $W_2$  and  $W_3$ , respectively, then, the sequences  $A_t$ ,  $C_t$ ,  $R_t$ ,  $Q_t$ , obtained in the four updating steps of Section 2.2 will converge respectively to finite limits  $A$ ,  $C$ ,  $R$  and  $Q$ . Also, writing  $A = (A_1, A_2)'$ , we will obtain

$$(1 - A_1)Q = 1 \quad (2.11)$$

$$A_2^2 Q = W_2 \quad (2.12)$$

$$(A_1^2 + A_1 A_2 - 2A_2)Q = W_1 - W_3 \quad (2.13)$$

and  $R$  will be given by

$$R = \begin{pmatrix} A_1 Q & A_2 Q \\ A_2 Q & A_1 A_2 Q - W_3 + W_2 \end{pmatrix}$$

In addition,  $A_1$  and  $A_2$  must satisfy the conditions

$$0 < A_1 < 1 \quad (2.14)$$

$$0 < A_2 < 4 - 2A_1 - 4(1 - A_1)^{1/2} < 2 \quad (2.15)$$

See West and Harrison (1989) for a proof.

We calculate the limit value of  $Q$  in the system (2.11) to (2.13) above. Let  $D = W_1 - W_3$ . From (2.13) we write

$$(A_1 - 2)A_2Q = D - A_1^2Q$$

and using (2.11) we have

$$(A_1 - 2)A_2Q = D - A_1(Q - 1)$$

and, from (2.11) again

$$(A_1 - 2)A_2Q = D + A_1 + 1 - Q$$

Multiply this by  $Q$  and use (2.11) to obtain

$$(A_1 - 2)A_2Q^2 = DQ + Q - 1 + Q - Q^2$$

Therefore

$$(2 - A_1)A_2Q^2 = Q^2 - (D + 2)Q + 1$$

and we use (2.11) again to get

$$(Q + 1)A_2Q = Q^2 - (D + 2)Q + 1$$

Squaring this last expression and using (2.12), we will get

$$(Q + 1)^2QW_2 = (Q^2 - (D + 2)Q + 1)^2$$

and we will finally arrive at

$$Q^4 - \alpha Q^3 + \beta Q^2 - \alpha Q + 1 = 0 \quad (2.16)$$

where

$$\alpha = W_2 + 2D + 4$$

$$\beta = D^2 + 4D + 6 - 2W_2$$

and (2.16) is an equation in  $Q$  of even degree  $d = 2j$  where the list of coefficients is symmetrical. Therefore, it can be reduced to an equation in  $\kappa^2$  of degree  $j$ , where  $\kappa = (Q + 1)/(Q - 1)$ . Using this transformation in (2.16) we obtain the equation

$$p\kappa^4 + q\kappa^2 + r = 0 \quad (2.17)$$

where

$$p = D^2 - 4W_2 \quad (2.18)$$

$$q = 4W_2 - 2D^2 - 8D \quad (2.19)$$

$$r = (D + 4)^2 \quad (2.20)$$

and we have  $\Delta = q^2 - 4pr = 16W_2(16 + 4D + W_2)$ . We want to verify that  $\Delta \geq 0$  and for this purpose it suffices to verify that  $4D + W_2 \geq 0$ . Observe, then, that

$$\begin{aligned} 4D + W_2 &= 4W_1 - 4W_3 + W_2 = \frac{4W_1^2 - 4W_1W_3 + W_1W_2}{W_1} \geq \\ &= \frac{4W_1^2 - 4W_1W_3 + W_3^2}{W_1} = \frac{(2W_1 - W_3)^2}{W_1} \geq 0 \end{aligned}$$

and, therefore, we have two real roots. From (2.20),  $r$  is always non-negative. Then, if  $p < 0$ , we guarantee there is a positive root. Suppose  $p > 0$  and we want to show the roots for (2.17) are positive, which is to show that  $q < 0$ . Using (2.18) to (2.20) we have  $p + q + r = 16$ . Then, we want to show that  $p + r > 16$ . Using (2.18) and (2.20), we get

$$p + r = 2D^2 + 8D - 4W_2 + 16$$

and we have to show, then, that  $D^2 + 4D > 2W_2$ , or, using (2.18),  $p + 4D > -2W_2$ . We have  $p + 4D > 4D$ , since  $p$  is positive, and  $-2W_3^2/W_1 \geq -2W_2$ . Then, it suffices to show that  $4D \geq -2W_3^2/W_1$ . But this is

$$4W_1^2 - 4W_3W_1 \geq -2W_3^2$$

which is trivially true, since it can be rewritten as

$$(2W_1 - W_3)^2 + W_3^2 \geq 0$$

It remains to show that we can always find a root which, in fact, is greater than one. For  $p < 0$ , this root will be  $-(q + \sqrt{\Delta})/2p$ . We verify this from the fact that  $p + q + r = 16$ . Then, we have

$$4p(p + q + r) < 0$$

and, from this,

$$q^2 - 4pr > q^2 + 4pq + 4p^2$$

which is  $\Delta > (q + 2p)^2$  and, from this, we have the desired result. For  $p > 0$ , we can show both roots are greater than one, by showing that  $r > p$ . From (2.18) and (2.20), this is equivalent to show that  $8D + 16 + 4W_2 > 0$ , or,  $2D + 4 + W_2 \geq 0$ . It suffices to show that  $2D + W_2 \geq 0$  and this is  $2W_1 + W_2 \geq 2W_3$ . But this is true since

$$(2W_1 + W_2)^2 \geq 4W_1W_2 \geq 4W_3^2$$

This shows (2.17) always has a real root which is greater than one. Now, we remember that  $A_1$  and  $A_2$  in the system (2.11) and (2.13) must satisfy (2.15):

$$0 < A_2 < 4 - 2A_1 - 4(1 - A_1)^{1/2}$$

From (2.11), we have

$$0 < A_2 < 4 - 2A_1 - 4Q^{-1/2}$$

Rewrite this as

$$0 < A_2 < 2 + 2(1 - A_1) - 4Q^{-1/2}$$

which is

$$0 < A_2 < 2 + 2Q^{-1} - 4Q^{-1/2} = 2(1 - Q^{-1/2})^2$$

Therefore

$$A_2^2 < 4(1 - Q^{-1/2})^4$$

and, we have

$$W_2 = QA_2^2 < 4(Q^{1/4})^4(1 - Q^{-1/2})^4 = 4(Q^{1/4} - Q^{-1/4})^4$$

which gives

$$W_2^{1/4} < \sqrt{2}(Q^{1/4} - Q^{-1/4})$$

Considering that there must be a unique solution to the system (2.11) to (2.13), and this is because the limit is unique, we must choose the biggest root of (2.16), since  $f(x) = x^{1/4} - x^{-1/4}$  is an increasing function of  $x$ . Now, because  $g(x) = (x + 1)/(x - 1)$  is a decreasing function of  $x$ , we must calculate  $Q$  from the smallest root in  $(1, \infty)$  of (2.17). Then, let  $E$  be this root and we finally arrive at

$$Q = \frac{\sqrt{E} + 1}{\sqrt{E} - 1}$$

and, now,  $A_1$  and  $A_2$  must be obtained from (2.11) and (2.12). Then, we have solved that system.

Finally, we must observe that the concept of discount factor can be extended here. We define the second order polynomial model with a single discount factor  $\delta$  as that model for which the variance of the evolution error is written as

$$W_t = J_2(1)C_{t-1}J_2(1)'(1 - \delta)/\delta$$

and this implies  $Var[\theta_t|D_t] = Var[\theta_t|D_{t-1}]/\delta$ , for all  $t$ . Thus, we have a model with a fixed multiplicative increase in uncertainty, the rate of decay in the information about the state vector being constant with time. From the practical point of view, a suitable approach employs a single constant discount factor for the trend and growth components. In this case, the limiting point predictor  $m_t$  converges in probability to that derived from exponentially weighted regression (Brown, 1962).

The same idea of a single, constant discount factor can be used when the model structure can be seen as of a single, canonical component. This will be the case, for example, if  $G$  is similar to an  $n \times n$  Jordan block. The idea of using a single discount approach is discussed in Brown (1962), Harrison (1965), Godolphin and Harrison (1975) and Harrison and Akram (1983).

## 2.5 Models with unknown observational variance

In many practical applications the observational variance  $V_t$  will be subject to uncertainty. When this is the case, the method we are using must include a learning procedure for the unknown observational variance. We discuss here the most simple case, namely,  $V_t$  being constant and unknown. More sophisticated situations can be found in West and Harrison (1989). We also refer to Smith and West (1983), and to West, Harrison and Pole (1987) for practical applications.

The key feature of our analysis is that we consider here a *scale-free* model, with all variances being scaled by the unknown observational variance  $V$ . That means, for example, that in the evolution equation (2.2) we have  $Var[\omega_t] = W_t = VW_t^*$ , for each  $t$ , where we assume previous knowledge, not of the sequence  $\{W_t\}$ , but, of the sequence  $\{W_t^*\}$ . Also, it will be convenient for our purposes to work with the *precision*  $\phi$ , defined by  $\phi = 1/V$ , rather than with  $V$  itself.

The extension of the more simple model for the case of unknown observational variance is as follows. We consider the observational and evolution equations

$$\begin{cases} y_t = F_t'\theta_t + v_t \\ \theta_t = G_t\theta_{t-1} + \omega_t, \end{cases}$$



where  $v_t \sim N[0, V_t]$  and  $\omega_t \sim N[0, VW_t^*]$ , the sequence of matrices  $\{W_t^*\}$  being completely known. Now, let  $\phi = 1/V$ . Then, our initial information consists of two prior distributions, namely

$$\begin{aligned}(\theta_0|D_0, \phi) &\sim N[m_0, VC_0^*] \\ (\phi|D_0) &\sim G[n_0/2, d_0/2]\end{aligned}$$

with the quantities  $m_0$ ,  $C_0^*$ ,  $n_0$  and  $d_0$  being pre-specified. Also, in the gamma prior distribution defined above, we observe that the prior mean is given by  $E[\phi|D_0] = n_0/d_0 = 1/S_0$ , where  $S_0$  is a prior point estimate of  $V$ . It is important to notice that we have

$$(d_0\phi|D_0) \sim \chi_{n_0}^2$$

In practice, to specify a prior distribution for  $\phi$ , we can choose a prior point estimate  $S_0$  of  $V$  and the associated number  $n_0$  of degrees of freedom.

The updating for the defined model is as follows. We depart from the information  $(\theta_{t-1}|D_{t-1}, V) \sim N[m_{t-1}, VC_{t-1}^*]$  and  $(\phi|D_{t-1}) \sim G[n_{t-1}/2, d_{t-1}/2]$ .

(a) First, we have, conditional on  $V$ ,

$$\begin{aligned}(\theta_t|D_{t-1}, V) &\sim N[a_t, VR_t^*] \\ (y_t|D_{t-1}, V) &\sim N[f_t, VQ_t^*] \\ (\theta_t|D_t, V) &\sim N[m_t, VC_t^*]\end{aligned}$$

where  $a_t = G_tm_{t-1}$ ,  $R_t^* = G_tC_{t-1}^*G_t' + W_t^*$ ,  $f_t = F_t'a_t$ ,  $Q_t^* = F_t'R_t^*F_t + 1$ , and the updating of the components is

$$\begin{aligned}m_t &= a_t + A_te_t \\ C_t^* &= R_t^* - A_tA_t'Q_t^*\end{aligned}$$

with  $A_t = R_t^*F_t/Q_t^*$  and  $e_t = y_t - f_t$ .

(b) The distribution of  $(\phi_t|D_t)$  can be obtained from Bayes' Theorem. We have

$$p(\phi|D_t) \propto p(\phi|D_{t-1})p(y_t|\phi, D_{t-1})$$

and this will give

$$(\phi|D_t) \sim G[n_t/2, d_t/2]$$

with  $n_t = n_{t-1} + 1$  and  $d_t = d_{t-1} + e_t^2/Q_t^*$ .

(c) Now, unconditional on  $V$ , the normal distributions in (a) will be replaced by  $T$  distributions. Let  $T_p[m, C]$  denote the multivariate  $T$  distribution of  $p$  degrees of

freedom with location  $m$  and scale  $C$ . We recall that if  $p > 1$ , then, the mean of the distribution exists and is equal to  $m$ . Also, if  $p > 2$ , the variance of the distribution exists and is equal to  $pC/(p-2)$ . The distribution approaches normality, as  $p$  tends to infinity, the limiting distribution being  $N[m, C]$ .

The unconditional distributions are given by

$$\begin{aligned}(\theta_{t-1}|D_{t-1}) &\sim T_{n_{t-1}}[m_{t-1}, C_{t-1}] \\(\theta_t|D_{t-1}) &\sim T_{n_{t-1}}[a_t, R_t] \\(y_t|D_{t-1}) &\sim T_{n_{t-1}}[f_t, Q_t] \\(\theta_t|D_t) &\sim T_{n_t}[m_t, C_t]\end{aligned}$$

where  $C_{t-1} = S_{t-1}C_{t-1}^*$ ,  $R_t = S_{t-1}R_t^*$ ,  $Q_t = S_{t-1}Q_t^*$  and  $C_t = S_tC_t^*$ , with  $S_{t-1} = d_{t-1}/n_{t-1}$  and  $S_t = d_t/n_t$ .

Some key points in the anterior updating must be commented. At time  $t$ , we have  $E[\phi|D_{t-1}] = 1/S_{t-1}$ , where  $S_{t-1} = d_{t-1}/n_{t-1}$ , a prior point estimate of  $V$ . Similarly, we have  $E[\phi|D_t] = 1/S_t$ . Observe that the updating equations for the  $T$  distributions in (c) are essentially the same equations that appear in (a), the variance  $V$  being substituted by its point estimate,  $S_{t-1}$  (or the updated estimate  $S_t$ , if we already have observed  $y_t$ ).

## 2.6 Reference Analysis of the DLM

### 2.6.1 Introduction

One of the key components of the standard DLM analysis is defined by the initial information we have about the state space  $\theta$ , this information being represented in the distribution  $(\theta_0|D_0)$ . The usual standard approach requires a proper prior distribution for initialization. It is important, then, to devise a way of obtaining a proper distribution for the state vector, when no prior knowledge is available at  $t = 0$ , other than the model itself, the only additional information being provided by the set of observations  $\{y_t\}$ . This *reference analysis* of the DLM, based on standard vague or uninformative priors (Bernardo, 1979), gives us, as the name suggests, a reference level, against which we can compare alternative approaches that will possibly use informative prior distributions.

In this section, we review the approach that is suggested by Pole and West (1989). It is also described in West and Harrison (1989), where proof of the results presented here can be found.

### 2.6.2 Updating equations

The aim here is to obtain a proper distribution for the state space using the information that is provided by the first observations of the process. We present here a Bayesian development. Departing from an initial reference prior at  $t = 1$ , we use a set of sequential updating equations, which is based on this reference prior and the information we collect with time. After each iteration, we expect to expand the subspace for which we have a proper distribution. This approach is performed until we have a non-singular covariance matrix for the state vector. From this point on, we have a proper prior, and can turn to standard analysis.

We present two cases. In the first case, the variance  $W_t$  of the evolution error is always non-singular. In the second case, the same variance is identically zero, the evolution equation being deterministic. In both cases we consider the updating equations when the variance  $V_t$  of the observational error is known, and when this variance is unknown but constant. It is assumed, for all the development here, that  $G_t$  is non-singular.

At time  $t = 1$ , we consider (see, e.g., Box and Tiao, 1973), when the observational variance is known, the reference prior

$$p(\theta_1|D_0) \propto \text{constant} \quad (2.21)$$

When the observational variance is equal to  $V$  and unknown, we consider the reference prior

$$p(\theta_1, V|D_0) \propto V^{-1} \quad (2.22)$$

CASE 1:  $W_t$  is non singular for all  $t$

For both models (known observational variance or constant and unknown observational variance) we sequentially define the following quantities:

$$P_t = G_t' W_t^{-1} G_t + K_{t-1} \quad (2.23)$$

$$H_t = W_t^{-1} - W_t^{-1} G_t P_t^{-1} G_t' W_t^{-1} \quad (2.24)$$

$$h_t = W_t^{-1} G_t P_t^{-1} k_{t-1} \quad (2.25)$$

where

$$K_t = \begin{cases} H_t + F_t F_t' & \text{if } V_t = V \text{ is unknown} \\ H_t + F_t F_t' / V_t & \text{if } V_t \text{ is known} \end{cases}$$

and

$$k_t = \begin{cases} h_t + F_t y_t & \text{if } V_t = V \text{ is unknown} \\ h_t + F_t y_t / V_t & \text{if } V_t \text{ is known} \end{cases}$$

with initial values  $H_1 = 0$  and  $h_1 = 0$ . Since we assume that  $G_t$  is non-singular, the matrix  $G'_t W_t^{-1} G_t$ , which appears in (2.23), must be positive definite. We also know that  $K_{t-1}$  is positive semidefinite, since it is the covariance matrix of  $(\theta_{t-1}|D_{t-1})$  (see equation 2.27). Hence,  $P_t$ , the sum of the two matrices in (2.23), has to be positive definite, and, consequently, non-singular. Therefore, the quantities in (2.24) and (2.25) are well defined.

If  $V_t$  is constant and unknown,  $W_t$  in the above equations must be replaced by  $W_t^*$ , and, in addition, define

$$\begin{aligned}\lambda_t &= \delta_{t-1} - k'_{t-1} P_t^{-1} k_{t-1} \\ \delta_t &= \lambda_t + y_t^2\end{aligned}$$

with initial value  $\lambda_1 = 0$ .

Suppose  $V_t$  is known. Applying Bayes' Theorem, we can verify, by induction that, using the reference prior (2.21), the prior and posterior distributions of the state vector at time  $t$  are given, respectively, by

$$p(\theta_t|D_{t-1}) \propto \exp\left\{-\frac{1}{2}(\theta'_t H_t \theta_t - 2\theta'_t h_t)\right\} \quad (2.26)$$

$$p(\theta_t|D_t) \propto \exp\left\{-\frac{1}{2}(\theta'_t K_t \theta_t - 2\theta'_t k_t)\right\} \quad (2.27)$$

If the variance  $V_t$  is constant and unknown, then, if we depart from the reference prior (2.22), we will have at time  $t$ , the prior and posterior joint distributions of the state vector and the variance  $V$  given, respectively, by

$$p(\theta_t, V|D_{t-1}) \propto V^{\frac{-(t+1)}{2}} \exp\left\{-\frac{1}{2}V^{-1}(\theta'_t H_t \theta_t - 2\theta'_t h_t + \lambda_t)\right\} \quad (2.28)$$

$$p(\theta_t, V|D_t) \propto V^{\frac{-(t+2)}{2}} \exp\left\{-\frac{1}{2}V^{-1}(\theta'_t K_t \theta_t - 2\theta'_t k_t + \delta_t)\right\} \quad (2.29)$$

It is readily seen that, in the case of variance known, after the distribution in (2.27) becomes proper, we turn to the standard updating, with

$$C_t = K_t^{-1} \text{ and } m_t = K_t^{-1} k_t$$

If the variance is unknown, then, after the distribution in (2.29) becomes proper, the posterior distributions of  $(\theta_t, V|D_t)$  are as in Section 2.5, with

$$C_t = S_t K_t^{-1} \text{ and } m_t = K_t^{-1} k_t$$

where  $S_t = d_t/n_t$ , as usual, with  $n_t = t-n$  and  $d_t = \delta_t - k'_t m_t$ ,  $n$  being the dimension of the state vector. If the distribution becomes proper after  $n$  steps, then  $n_{n+1} = 1$  and it is easily shown that  $d_{t+1} = S_{n+1} = e_{n+1}^2/Q_{n+1}^*$ .

## CASE 2: $W_t = 0$

For this case, let  $M_t = G_t^{-1}$ , and define recursively the quantities

$$H_t = M_t' K_{t-1} M_t$$

$$h_t = M_t' k_{t-1}$$

$$\lambda_t = \delta_{t-1}$$

Using again the Bayes' Theorem we can show by induction that the prior and posterior distributions of  $\theta_t$  and  $V$  will again be of the forms (2.26) and (2.27) if  $V$  is known, and of the forms (2.28) and (2.29) if  $V$  is constant and unknown.

### 2.7 Dynamic Generalised Linear Models

#### 2.7.1 Motivation

We turn now to the extension of the DLM to non-normal observations. In many cases, when treating with some skewed distributions, we can find a suitable transformation such that we can apply the standard DLM to the transformed data. However, the interpretation of the parameters that will be used in the model for this transformed observations will not always be easy, and this can cause a bit of confusion for the modeller. It becomes necessary, therefore, to have a model that can be directly applied to the original observations. This will be particularly true when the original data is in the form of counts, especially if we are dealing with integer numbers of small magnitude. For these cases, any transformation trying to achieve normality will often be nonsense. For these reasons, we are motivated to extend the ideas of the DLM, arriving at the exponential family models. Such models are discussed in this section, the primary references being Migon and Harrison (1985), West and Harrison (1986), West, Harrison and Migon (1985).

#### 2.7.2 Exponential family distribution

The most important extension of the normal DLM to non-normal observations that have been studied has its starting point in the framework of the Generalised Linear Models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983). This extension considers that the distribution of the observations falls into a particular class which is called the *exponential family*. This class is defined as follows. Let  $y_t$  represent, as usual, the observations of the process we are studying. The density function (or probability function, if it is discrete) for the distribution of  $y_t$  belongs to the exponential family if and only if it can be expressed in the form

$$p(y_t|\eta_t, V_t) = b(y_t, V_t) \exp\{V_t^{-1}[y_t\eta_t - a(\eta_t)]\} \quad (2.30)$$

where  $V_t > 0$  and  $\eta_t$  are real quantities that are referred to, respectively, as the *scale parameter* and the *natural parameter* of the distribution, and  $a(\cdot)$  is a twice differentiable convex function.

Let  $R_t(s) = \log E[\exp(sy_t)|\eta_t, V_t]$ , the logarithm of the moment generating function of  $(y_t|\eta_t, V_t)$ . The mean and variance of  $(y_t|\eta_t, V_t)$  can be calculated using the well known result that states that  $E[y_t|\eta_t, V_t] = R'_t(0)$  and  $Var[y_t|\eta_t, V_t] = R''_t(0)$ , the derivatives being calculated with respect to  $s$ . From (2.30) we can easily derive

$$R_t(s) = V_t^{-1}[a(\eta_t + sV_t) - a(\eta_t)]$$

and we have

$$R'_t(s) = \frac{d}{ds}a(\eta_t + sV_t)$$

$$R''_t(s) = V_t \frac{d^2}{ds^2}a(\eta_t + sV_t)$$

From the expressions above we get

$$E[y_t|\eta_t, V_t] = R'_t(0) = \mu_t = \dot{a}(\eta_t) \quad (2.31)$$

$$Var[y_t|\eta_t, V_t] = R''_t(0) = V_t \ddot{a}(\eta_t) \quad (2.32)$$

where the dots now represent the derivative with respect to  $\eta_t$ . The first two derivatives of  $a(\cdot)$  are, for obvious reasons, respectively called the *mean function* and the *variance function* of the distribution. From (2.32) it is readily seen that we necessarily must have positivity of the variance function. Hence, the mean function must be monotonically increasing and (2.31) defines a bijection between the mean and the natural parameter of the distribution.

### 2.7.3 Definition of DGLM

We now define the DGLM (dynamic generalised linear model), which extends the DLM concept for the case of non-normal observations. Let  $y_t$  be the process under study and suppose that the density  $p(y_t|\eta_t)$  belongs to the exponential family, being given by (2.30). Here, we assume that the scale parameter  $V_t$  is known for all  $t$ , and the explicit dependence on  $V_t$  is dropped. We consider a parameter  $\lambda_t$  which is related to the natural parameter  $\eta_t$  via the equation

$$\lambda_t = g(\eta_t) \quad (2.33)$$

where  $g(\cdot)$  is a continuous monotonic function mapping  $\eta_t$  to the real line. This parameter  $\lambda_t$ , a transformation of  $\eta_t$ , is related to a quantity  $\theta_t$  via a time dependent linear function of the form

$$\lambda_t = F'_t \theta_t \quad (2.34)$$

Here,  $F_t$  is an  $n$ -dimensional regression vector, known for every instant  $t$  and  $\theta_t$  is the  $n$ -dimensional vector of parameters. This state vector has its evolution governed by the equation

$$\theta_t = G_t \theta_{t-1} + \omega_t \quad (2.35)$$

where  $G_t$  is an  $n \times n$  evolution matrix, also known for all  $t$ , and  $\omega_t$  is a zero mean uncorrelated sequence, with  $\omega_t$  uncorrelated with  $\theta_{t-1}$ . We also assume that  $W_t = \text{Var}[\omega_t]$  is known for all  $t$ . Then, our observation model is defined by (2.30) together with (2.33) and (2.34) while the evolution model is given by (2.35). The latter is identical to (2.2), except by the fact the we do not necessarily require normality of distributions.

This definition extends the observation model for the DLM, the additional component  $g(\cdot)$  providing a link between the linear regression in (2.34) and the observational distribution (2.30). The standard DLM is here the particular case for which the distribution in (2.30) is  $N(\lambda_t, V_t)$ , the distributions of  $\theta_t$  and  $\omega_t$  are normal distributions and  $g(\cdot)$  is the identity mapping.

#### 2.7.4 Updating of DGLM

Our aim is to develop a sequential updating procedure for the GDLM, as was done for the standard DLM. In order to do so, we first consider a reformulation of the standard sequential procedure for the DLM, and then, extend it to the GDLM. Therefore, we reformulate the updating that was described in Section 2.2.2 as follows

Step 0: At time  $t-1$  we are provided with the posterior distribution  $(\theta_{t-1}|D_{t-1}) \sim N[m_{t-1}, C_{t-1}]$ . From this posterior distribution and the evolution equation we can calculate the prior for  $\theta_t$  at time  $t$ , obtaining

$$(\theta_t|D_{t-1}) \sim N[a_t, R_t]$$

where  $a_t = G_t m_{t-1}$  and  $R_t = G_t C_{t-1} G_t' + W_t$ . Up to here, we still proceed in the same way we did before. Now, we consider the next step

Step 1: We work with  $\mu_t = E[y_t|\eta_t] = F_t' \theta_t$ . The joint prior distribution for  $\mu_t$  and  $\theta_t$  is given by:

$$\begin{pmatrix} \mu_t \\ \theta_t \end{pmatrix} \Big| D_{t-1} \sim N \left[ \begin{pmatrix} f_t \\ a_t \end{pmatrix}, \begin{pmatrix} q_t & F_t' R_t \\ R_t F_t & R_t \end{pmatrix} \right] \quad (2.36)$$

where  $f_t = F_t' a_t$  and  $q_t = F_t' R_t F_t$ .

Step 2: It is possible now to calculate the one-step ahead forecasting for  $y_t$ . We know that  $(y_t|\mu_t) \sim N[\mu_t, V_t]$  and we have the prior  $(\mu_t|D_{t-1}) \sim N[f_t, q_t]$ . The one-step ahead forecast distribution is obtained as

$$p(y_t|D_{t-1}) = \int p(y_t|\mu_t) p(\mu_t|D_{t-1}) d\mu_t$$

and this will give us  $(y_t|D_{t-1}) \sim N[f_t, Q_t]$ , where  $Q_t = q_t + V_t$ .

Step 3: Now we obtain the posterior for  $\mu_t$ . This can be done by the use of Bayes' Theorem

$$p(\mu_t|D_t) \propto p(\mu_t|D_{t-1})p(y_t|\mu_t)$$

and we will get  $(\mu_t|D_t) \sim N[f_t^*, q_t^*]$ , where

$$f_t^* = f_t + q_t(y_t - f_t)/Q_t \quad (2.37)$$

$$q_t^* = q_t - q_t^2/Q_t \quad (2.38)$$

Step 4: We are now in position to calculate the posterior distribution for  $\theta_t$ , closing the updating procedure. We consider, therefore, the joint posterior for  $\mu_t$  and  $\theta_t$ , applying Bayes' Theorem, to obtain

$$\begin{aligned} p(\mu_t, \theta_t|D_t) &\propto p(\mu_t, \theta_t|D_{t-1})p(y_t|\mu_t) \\ &= [p(\theta_t|\mu_t, D_{t-1})p(\mu_t|D_{t-1})]p(y_t|\mu_t) \\ &= p(\theta_t|\mu_t, D_{t-1})[p(\mu_t|D_{t-1})p(y_t|\mu_t)] \\ &\propto p(\theta_t|\mu_t, D_{t-1})p(\mu_t|D_t) \end{aligned}$$

From above, we can see that  $\theta_t$  is conditionally independent of  $y_t$ , given  $\mu_t$ . The posterior distribution for  $\theta_t$  will, then, be obtained as

$$p(\theta_t|D_t) = \int p(\theta_t|\mu_t, D_{t-1})p(\mu_t|D_t) d\mu_t$$

It is important to observe that the information that  $y_t$  brings about  $\theta_t$  is reflected in the anterior expression through the posterior  $(\mu_t|D_t)$ . The first probability density in the integrand is that of the conditional normal distribution  $(\theta_t|\mu_t, D_{t-1})$ . From (2.36) it is readily seen to be given by

$$(\theta_t|\mu_t, D_{t-1}) \sim N[a_t + R_t F_t(\mu_t - f_t)/q_t, R_t - R_t F_t R_t'/q_t]$$

Now the moments for  $(\theta_t|D_t)$  can be calculated. We have

$$\begin{aligned} m_t &= E[\theta_t|D_t] \\ &= E[E[\theta_t|\mu_t, D_{t-1}]|D_t] \\ &= E[a_t + R_t F_t(\mu_t - f_t)/q_t|D_t] \\ &= a_t + R_t F_t(f_t^* - f_t)/q_t \end{aligned} \quad (2.39)$$



and

$$\begin{aligned}
C_t &= \text{Var}[\theta_t | D_t] \\
&= E[\text{Var}[\theta_t | \mu_t, D_{t-1}] | D_t] + \text{Var}[E[\theta_t | \mu_t, D_{t-1}] | D_t] \\
&= E[R_t - R_t F_t F_t' R_t / q_t] + \text{Var}[a_t + R_t F_t (\mu_t - f_t) / q_t | D_t] \\
&= R_t - R_t F_t F_t' R_t / q_t + R_t F_t F_t' R_t q_t^* / q_t^2 \\
&= R_t - R_t F_t F_t' R_t (1 - q_t^* / q_t) / q_t
\end{aligned} \tag{2.40}$$

Substituting (2.37) and (2.38) in (2.39) and (2.40) we obtain the usual expressions  $m_t = a_t + R_t F_t (y_t - F_t' a_t) / Q_t$  and  $C_t = R_t - R_t F_t F_t' R_t / Q_t$ . Then, we have completed the update.

The alternative development we have just presented for obtaining the updating equations in the DLM can now be extended to the non-normal case. Although there is no general, exact analysis for the DGLM (since the observations are not necessarily normal and the mean of their distribution is a function not necessarily linear of the state vector), it is possible to develop an approximation, paralleling the steps above. In this development we drop the normality assumption for the state vector. We consider, then, that, at time  $t - 1$ , we have a posterior distribution  $(\theta_{t-1} | D_{t-1})$  which is only partially specified in terms of its first two moments. Similarly, the distribution of the error in the evolution equation is now only partially specified by its first two moments. We use the notation  $X \sim [\mu, V]$  meaning that the random variable  $X$  has mean  $\mu$  and variance  $V$ . Then, we have

$$(\theta_{t-1} | D_{t-1}) \sim [m_{t-1}, C_{t-1}]$$

and, from the independence assumption, we will have

$$(\theta_t | D_{t-1}) \sim [a_t, R_t] \tag{2.41}$$

where  $a_t = G_t m_{t-1}$  and  $R_t = G_t C_{t-1} G_t' + W_t$ . From this point, we parallel the four steps described anteriorly. We reiterate the notation we have been using. We consider  $\mu_t = E[y_t | \eta_t]$ , where  $\eta_t$  is the natural parameter of the distribution for  $y_t$ , and we have  $\mu_t = \dot{a}(\eta_t)$ . Also, the natural parameter  $\eta_t$  is linked to the parameter  $\lambda_t$  via  $\lambda_t = g(\eta_t)$ . The parameter  $\lambda_t$ , on its turn, relates to the state vector  $\theta_t$  via the regression equation  $\lambda_t = F_t' \theta_t$ . Because  $\dot{a}(\cdot)$  and  $g(\cdot)$  are real bijective functions, we are allowed to work with  $\mu_t$ ,  $\eta_t$  or  $\lambda_t$ , interchangeably.

Step 1: From (2.41), we can readily obtain the first two moments of the joint prior distribution for  $\lambda_t$  and  $\theta_t$ , as

$$\begin{pmatrix} \lambda_t \\ \theta_t \end{pmatrix} \Big| D_{t-1} \sim \left[ \begin{pmatrix} f_t \\ a_t \end{pmatrix}, \begin{pmatrix} q_t & F_t' R_t \\ R_t F_t & R_t \end{pmatrix} \right]$$

where  $f_t = F_t' a_t$  and  $q_t = F_t' R_t F_t$ , and we still work as in the normal case.

Step 2: We are now interested in the one-step ahead forecasting for  $y_t$ . From (2.30), the relevant information needed for forecasting  $y_t$  is contained in the distribution of  $(\eta_t|D_{t-1})$ . But, now, this distribution is, in principle, only partially specified, since we do not necessarily know the full distributional form of  $\lambda_t = g(\eta_t)$ , but only the first two moments of its distribution. Therefore, to obtain the forecast we want, we need to consider other assumptions about the prior distribution for  $\eta_t$ . One reasonable alternative is to work with a conjugate prior for  $\eta_t$ ; the prior must be consistent with the mean and variance we have for  $\lambda_t$ . From (2.30), a prior density for  $\eta_t$  must have the form

$$p(\eta_t|D_{t-1}) = c(r_t, s_t) \exp[r_t \eta_t - s_t a(\eta_t)] \quad (2.42)$$

where  $r_t$  and  $s_t$  are the parameters of the distribution, that must be consistent with the moments for  $\lambda_t$ . That means,  $r_t$  and  $s_t$  must be such that

$$\begin{aligned} E[g(\eta_t)|D_{t-1}] &= f_t \\ Var[g(\eta_t)|D_{t-1}] &= q_t \end{aligned}$$

The one-step ahead forecast distribution can, then, be calculated via

$$p(y_t|D_{t-1}) = \int p(y_t|\eta_t) p(\eta_t|D_{t-1}) d\eta_t$$

with the two densities given respectively by (2.30) and (2.42). We, then, have

$$p(y_t|D_{t-1}) = \frac{c(r_t, s_t) b(y_t, V_t)}{c(r_t + V_t^{-1} y_t, s_t + V_t^{-1})}$$

Step 3: It is immediate to obtain the posterior distribution for  $\eta_t$ ; this is given by

$$p(\eta_t|D_t) = c(r_t^*, s_t^*) \exp[r_t^* \eta_t - s_t^* a(\eta_t)] \quad (2.43)$$

with  $r_t^* = r_t + V_t^{-1} y_t$  and  $s_t^* = s_t + V_t^{-1}$ . It is important, then, to observe that (2.42) is, indeed, a conjugate prior distribution for (2.30). We can now calculate the posterior moments for  $\lambda_t = g(\eta_t)$ . We denote

$$\begin{aligned} f_t^* &= E[g(\eta_t)|D_t] \\ q_t^* &= Var[g(\eta_t)|D_t] \end{aligned}$$

Step 4: We want now to obtain the posterior moments for  $(\theta_t|D_t)$ . Again, we depart from the joint posterior distribution for  $\lambda_t$  and  $\theta_t$ .

$$\begin{aligned} p(\lambda_t, \theta_t|D_t) &\propto p(\lambda_t, \theta_t|D_{t-1}) p(y_t|\lambda_t) \\ &= [p(\theta_t|\lambda_t, D_{t-1}) p(\lambda_t|D_{t-1})] p(y_t|\lambda_t) \\ &= p(\theta_t|\lambda_t, D_{t-1}) [p(\lambda_t|D_{t-1}) p(y_t|\lambda_t)] \\ &\propto p(\theta_t|\lambda_t, D_{t-1}) p(\lambda_t|D_t) \end{aligned}$$

Again,  $\theta_t$  is conditionally independent of  $y_t$ , given  $\lambda_t$ . The posterior distribution for  $\theta_t$  is

$$p(\theta_t|D_t) = \int p(\theta_t|\lambda_t, D_{t-1})p(\lambda_t|D_t) d\lambda_t$$

As in the anterior case, the information about  $\theta_t$  that is brought by  $y_t$  is used in the above integral through  $p(\lambda_t|D_t)$ . The last probability density in the integrand can be readily calculated since we have (2.43). The first density is not always fully specified, but we observe that this is not strictly necessary for our purposes, since we do not want the full posterior distribution for  $(\theta_t|D_t)$ , but just its first two moments. Therefore, we only need the first two moments of the distribution of  $(\theta_t|\lambda_t, D_{t-1})$  to complete the updating. These moments cannot always be calculated, but they can be *estimated* from standard Bayesian techniques. We recall the general result for linear Bayes estimation. Further discussion can be found in Hartigan (1969) and Goldstein (1976). Suppose we want to use the information of an observation  $Y$  to estimate a parameter  $\theta$ , through a function  $d(Y)$  of this observation. The linear function  $d(Y) = h + HY$  of  $Y$  that minimizes the overall risk

$$r(d) = \text{trace } E[(\theta - d)(\theta - d)'] \quad (2.44)$$

is that defined by  $d^*(Y) = h^* + H^*Y$ , where

$$h^* = a - SQ^{-1}f \quad (2.45)$$

$$H^* = SQ^{-1} \quad (2.46)$$

if the joint distribution of  $Y$  and  $\theta$  is partially specified by its first two moments as

$$\begin{pmatrix} Y \\ \theta \end{pmatrix} \sim \left[ \begin{pmatrix} f \\ a \end{pmatrix}, \begin{pmatrix} Q & S' \\ S & R \end{pmatrix} \right]$$

The estimator  $d^*(Y) = a + SQ^{-1}(Y - f)$  is called the linear Bayes' estimator (LBE) of  $\theta$  based upon  $Y$ . The value of the overall risk (2.44) at  $d = d^*$  is  $r(d^*) = \text{trace } (C)$ , where

$$C = R - SQ^{-1}S'$$

In our particular case, we have that prior joint distribution of Step 1. From that, we estimate the conditional mean  $E[\theta_t|\lambda_t, D_{t-1}]$  (which minimizes the expected quadratic loss  $E[\text{trace } (\theta_t - d(\lambda_t))(\theta_t - d(\lambda_t))'|\lambda_t]$  between all estimators of  $\theta_t$  based on  $\lambda_t$ ). Using (2.45) and (2.46), we have the LBE optimal estimate of this conditional mean given by

$$\hat{E} = a_t + R_t F_t (\lambda_t - f_t) / q_t \quad (2.47)$$

for all  $\lambda_t$ . West and Harrison (1988), p. 561, suggest that  $Var[\theta_t|\lambda_t, D_{t-1}]$  be estimated by the formula

$$\hat{V} = R_t - R_t F_t F_t' R_t / q_t \quad (2.48)$$

We now use the expressions for the moments of the marginal distribution in terms of the conditional moments

$$\begin{aligned} E[\theta_t|D_t] &= E[E\{\theta_t|\lambda_t, D_{t-1}\}|D_t] \\ Var[\theta_t|D_t] &= Var[E\{\theta_t|\lambda_t, D_{t-1}\}|D_t] + E[Var\{\theta_t|\lambda_t, D_{t-1}\}|D_t] \end{aligned}$$

and substitute the estimates given by (2.47) and (2.48), to obtain the estimated posterior moments, as

$$\begin{aligned} m_t &= E[\hat{E}] \\ &= E[a_t + R_t F_t(\lambda_t - f_t)/q_t|D_t] \\ &= a_t + R_t F_t(f_t^* - f_t)/q_t \end{aligned} \quad (2.49)$$

and

$$\begin{aligned} C_t &= E[\hat{V}] + Var[\hat{E}] \\ &= E[R_t - R_t F_t F_t' R_t / q_t] + Var[a_t + R_t F_t(\lambda_t - f_t)/q_t|D_t] \\ &= R_t - R_t F_t F_t' R_t / q_t + R_t F_t F_t' R_t q_t^* / q_t^2 \\ &= R_t - R_t F_t F_t' R_t (1 - q_t^* / q_t) / q_t \end{aligned} \quad (2.50)$$

Substituting the values of  $f_t^*$  and  $q_t^*$  obtained in Step 3 completes the updating.

We can now obtain the  $k$  steps ahead forecast at time  $t$ . Using the posterior moments obtained at time  $t$ , we can make the one step analysis.

## 2.8 Related work

Much theoretical and practical work related to the material presented here has been done in many fields of research. Classic textbooks in Bayesian modelling and forecasting include Box and Tiao (1973), Zellner (1971), Aitchison and Dunsmore (1975). On dynamic and sequential modelling the books by Astrom (1970) and Young (1984) are of note.

In discounting, the books by Brown (1959, 1963) constitute a classic reference and we should also mention the works of Morrison (1969), Godolphin and Harrison (1975), Harrison and Akram (1983), Ameen and Harrison (1985). Polynomial forecast functions have been studied in Harrison (1965, 1967), Godolphin and Harrison

(1975), Godolphin and Stone (1980). Procedures for variance learning with practical applications can be found in Smith and West (1983), West, Harrison and Pole (1987).

Several applications and illustrations of the DGLM can be seen in the literature. See, for example, the study in advertising awareness in Migon and Harrison (1985). Many specific cases of the DGLM are discussed in West and Harrison (1986) and also in West, Harrison and Migon (1985). Also, related models are considered from different viewpoints by Azzalini (1983), Smith (1979, 1988), Souza (1981) and Smith and Miller (1986).

There has been considerable non Bayesian work on dynamic modelling and forecasting, particularly in control engineering. The updating algorithm of Kalman (1960) has been widely used in connection with state space models. Anderson and Moore (1971) and Jazwinski (1970) are examples of good textbooks. Statisticians, econometricians and others have been worked on related fields, as, for example, Akaike (1974), Duncan and Horne (1972), Harvey (1981), Theil (1981).

### CHAPTER 3

## REFERENCE ANALYSIS OF THE DLM

When applying the classical updating procedure for the Dynamic Linear Model, we need to begin with prior information about the parameter vector,  $\theta$ . A common procedure, which is a basic idea on reference analysis, consists in obtaining this initial information, making use of the observations at the beginning of data. However, the postulated model for data evolution can also supply initial information, at least, concerning a subspace of the state space. That means we can assign a proper distribution for a certain subspace of the state, provided we know the covariance matrix of the errors in the system equation. The method we develop here departs from that of Pole and West (1989). It uses the first observations much more efficiently in order to enlarge the subspace for which we have a proper distribution. This subspace is enlarged until we get a proper distribution for the entire state space. Once this point is reached, we may turn to the common updating algorithm of the DLM.

We review the result of Section 2.6. In the sequential updating algorithm of Pole and West, the prior and posterior covariance matrices of the state vector at time  $t$  are given, respectively, by  $H_t$  and  $K_t$ . These matrices are recursively obtained from the equations

$$P_t = G_t' W_t^{-1} G_t + K_{t-1} \quad (3.1)$$

$$H_t = W_t^{-1} - W_t^{-1} G_t P_t^{-1} G_t' W_t^{-1} \quad (3.2)$$

where

$$K_t = \begin{cases} H_t + F_t F_t' & \text{if } V_t = V \text{ is unknown} \\ H_t + F_t F_t' / V_t & \text{if } V_t \text{ is known} \end{cases}$$

with initial value  $H_1 = 0$ . Here  $\{F_t, G_t, V_t, W_t\}$  correspond to the usual parameters defining the DLM under study. Since  $G_t$  is supposed to be non-singular, we can guarantee that  $P_t$ , defined by (3.1) is also.

Let  $n$  be the dimension of the state space. We will verify that, in the above equations,  $K_t$  will be non-singular, only after at least  $n$  updatings and that this depends fundamentally on the vectors  $\{F_t\}$ . Given a matrix  $A$ , denote its rank by  $r(A)$  and let  $\ker(A)$  be the subspace  $\{x \in R^n | Ax = 0\}$ . Then, from (3.2), we have

$$r(H_t) = r(I - G_t P_t^{-1} G_t' W_t^{-1}) \quad (3.3)$$

Now, we know that

$$n - r(H_t) = n - r(I - G_t P_t^{-1} G_t' W_t^{-1}) = \dim\{\ker(I - G_t P_t^{-1} G_t' W_t^{-1})\}$$

## CHAPTER 3

### REFERENCE ANALYSIS OF THE DLM

When applying the classical updating procedure for the Dynamic Linear Model, we need to begin with prior information about the parameter vector,  $\theta$ . A common procedure, which is a basic idea on reference analysis, consists in obtaining this initial information, making use of the observations at the beginning of data. However, the postulated model for data evolution can also supply initial information, at least, concerning a subspace of the state space. That means we can assign a proper distribution for a certain subspace of the state, provided we know the covariance matrix of the errors in the system equation. The method we develop here departs from that of Pole and West (1989). It uses the first observations much more efficiently in order to enlarge the subspace for which we have a proper distribution. This subspace is enlarged until we get a proper distribution for the entire state space. Once this point is reached, we may turn to the common updating algorithm of the DLM.

We review the result of Section 2.6. In the sequential updating algorithm of Pole and West, the prior and posterior covariance matrices of the state vector at time  $t$  are given, respectively, by  $H_t$  and  $K_t$ . These matrices are recursively obtained from the equations

$$P_t = G_t' W_t^{-1} G_t + K_{t-1} \quad (3.1)$$

$$H_t = W_t^{-1} - W_t^{-1} G_t P_t^{-1} G_t' W_t^{-1} \quad (3.2)$$

where

$$K_t = \begin{cases} H_t + F_t F_t' & \text{if } V_t = V \text{ is unknown} \\ H_t + F_t F_t' / V_t & \text{if } V_t \text{ is known} \end{cases}$$

with initial value  $H_1 = 0$ . Here  $\{F_t, G_t, V_t, W_t\}$  correspond to the usual parameters defining the DLM under study. Since  $G_t$  is supposed to be non-singular, we can guarantee that  $P_t$ , defined by (3.1) is also.

Let  $n$  be the dimension of the state space. We will verify that, in the above equations,  $K_t$  will be non-singular, only after at least  $n$  updatings and that this depends fundamentally on the vectors  $\{F_t\}$ . Given a matrix  $A$ , denote its rank by  $r(A)$  and let  $\ker(A)$  be the subspace  $\{x \in R^n | Ax = 0\}$ . Then, from (3.2), we have

$$r(H_t) = r(I - G_t P_t^{-1} G_t' W_t^{-1}) \quad (3.3)$$

Now, we know that

$$n - r(H_t) = n - r(I - G_t P_t^{-1} G_t' W_t^{-1}) = \dim\{\ker(I - G_t P_t^{-1} G_t' W_t^{-1})\}$$

But we have

$$x \in \ker(I - G_t P_t^{-1} G_t' W_t^{-1}) \iff x = G_t P_t^{-1} G_t' W_t^{-1} x \iff P_t G_t^{-1} x = G_t' W_t^{-1} x \quad (3.4)$$

From (3.1), we have

$$P_t G_t^{-1} x = G_t' W_t^{-1} x \iff (G_t' W_t^{-1} G_t + K_{t-1}) G_t^{-1} x = G_t' W_t^{-1} x \quad (3.5)$$

From (3.4) and (3.5),

$$x \in \ker(I - G_t P_t^{-1} G_t' W_t^{-1}) \iff K_{t-1} G_t^{-1} x = 0 \iff x \in \ker(K_{t-1} G_t^{-1})$$

Hence

$$r(H_t) = r(I - G_t P_t^{-1} G_t' W_t^{-1}) = r(K_{t-1} G_t^{-1}) = r(K_{t-1})$$

which means

$$r(H_{t+1}) = r(H_t + F_t F_t') \quad (3.6)$$

From (3.6), it is clear that in the best situation, the rank improves as  $r(H_{t+1}) = r(H_t) + 1$ , and this is the case where  $F_t$  does not belong to the space spanned by the columns of  $H_t$ . As  $H_1 = 0$ , we need at least  $n$  iterations to get a full rank covariance matrix and a proper prior.

We can readily see, then, from (3.6), that, to arrive at a proper prior, we depend fundamentally on the behaviour of  $\{F_t\}$ . We can see, for example, that there may exist situations when we cannot obtain a full proper prior, for example, if  $F$  becomes constant before we can get a full rank matrix  $H_t$ . However, this will be of no importance if we want to make forecasts for that particular subspace of the state space for which a proper prior exists.

The big advantage of the approach we propose is that useful forecasts can be made well before a full proper distribution is obtained. Hence, for example, if  $F$  is maintained constant over the first, let's say, one hundred observations, no information on the variation of  $F$  will be available from those observations and a proper distribution for  $\theta_t$  will not be achieved. However, if  $F$  is to remain at this same value, this does not constitute a problem, since we are only interested in a proper distribution conditional on  $F$  remaining at this same value. Essentially, the idea is that after a certain iteration, there will always be proper conditional distributions, and that, if these conditions are satisfied over the forecasting horizon, then a proper conditional forecast distribution is available.

The procedure of Pole and West(1989) does not deal with this, not with the fact that sometimes an initial proper distribution can be obtained, at least for a certain



subspace of the state. Giving an example, suppose  $y_t$  represents an AR(1) process, with evolution in time given by the expression:

$$y_t = \rho y_{t-1} + a_t$$

where  $a_t$  is a white noise with known variance  $\sigma_a^2$ ,  $|\rho| < 1$ , and there is a finite  $K$  such that  $E[y_t^2] < K$ , for all  $t$ . In this case,  $y_t$  will be the limit in squared mean, given below:

$$y_t = \sum_{i=0}^{\infty} \rho^i a_{t-i}$$

which is a zero mean stationary process with variance given by

$$E[y_t^2] = \sigma_a^2 \sum_{i=0}^{\infty} \rho^{2i} = \sigma_a^2 / (1 - \rho^2)$$

Hence, we can propose the initial prior mean and variance of  $y_t$ , based on the mean and variance of  $a_t$ .

Similarly, consider the vector process defined by the equation:

$$\theta_t = G\theta_{t-1} + \omega_t \quad (3.7)$$

where  $G$  is a constant system matrix and the  $\omega_t$  are uncorrelated (0,W) random vectors. If all eigenvalues of  $G$  lie inside the unit circle, then  $\theta_t$  can be expressed as an infinite moving average of the  $\omega_t$ , in the form  $\theta_t = \sum_{j=0}^{\infty} G^j \omega_{t-j}$ . Hence, if the W matrix is known, we can easily obtain the covariance matrix for  $\theta_t$ .

Now, suppose we can write (3.7) in the partitioned form:

$$\begin{bmatrix} \theta_{1,t} \\ \theta_{2,t} \end{bmatrix} = \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix} \begin{bmatrix} \theta_{1,t-1} \\ \theta_{2,t-1} \end{bmatrix} + \begin{bmatrix} \omega_{1,t} \\ \omega_{2,t} \end{bmatrix}$$

If, for example,  $G_1$  is a square matrix with all eigenvalues inside the unit circle, then, we can obtain an initial information concerning a subspace of the state space vector, i.e., we can derive mean and variance for  $\theta_1 = L\theta$ , where  $L = [G_1 \ 0]$ . In addition, consider the observational equation:

$$y_t = F_t' \theta_t + v_t$$

If  $F_t'$  belongs to the row space of  $L$ , we may calculate the first two moments for future values of  $y$ .

In the general case, let's consider the DLM model given by its observational and system equation:

$$\begin{cases} y_t = F_t' \theta_t + v_t & (3.8.1) \end{cases}$$

$$\begin{cases} \theta_t = G_t \theta_{t-1} + \omega_t & (3.8.2) \end{cases}$$

If  $G_t$  is a constant matrix  $G$  with (not necessarily all) eigenvalues inside the unit circle, then our model supplies an initial information concerning a subspace of the state space. That means it provides information about some  $\theta^* = L\theta$ , where  $L$  is obtained by choosing some appropriate rows of  $G$ . Thus, we have an initial prior information ( $\theta_1^*|D_0$ ). Now, if the regression vector  $F_1'$  belongs to the row space of  $L$ , we can forecast the future value of the observed variable, since  $E[y_1]$  can be expressed as a linear combination of the components of  $E[\theta_1^*]$ . If this is not the case, we can then use the information supplied by  $y_1$  (which is linked to  $F_1'$  through the equation 3.8.2) in order to enlarge our “information subspace”. Hence, after a finite number of steps, we will get a proper distribution for the state vector.

Consider the DLM defined by the system (3.8). We suppose that  $v_t$  and  $\omega_t$  are white noises such that  $E[v_t] = 0$ ,  $E[\omega_t] = 0$ ,  $E[v_t\omega_s] = 0$ , and  $E[\theta_t'\omega_{t+k}] = 0, \forall t, s$  and  $\forall k > 0$ . We also suppose that the variance  $V_t$  of  $v_t$  and the covariance matrix  $W_t$  of  $\omega_t$  are known, except for a scale factor  $\phi$ , such that  $V_t = V_t^*/\phi$  and  $W_t = W_t^*/\phi$ , where  $V_t^*$  and  $W_t^*$  are known (we will assume  $V_t^* \equiv 1$ , for simplicity). The distribution of all our variables in (3.8), conditioned on  $\phi$ , will be considered to be normal.

At time  $t$  we have prior information about  $\theta_t$  and  $\phi$ , based upon the model and all data until  $t - 1$ , such that

$$(L_t\theta_t|\phi, D_{t-1}) \sim N[a_t; (\phi P_t)^{-1}]$$

$$(\phi|D_{t-1}) \sim G(n_{t-1}/2; n_{t-1}S_{t-1}/2)$$

Here,  $L_t$  is a full rank  $r_t \times n$  matrix,  $r_t < n$ , where  $n$  is the dimension of the state space vector ( $L_1$  will be a  $1 \times n$  zero matrix if there is no initial information).

Let  $R_t$  be the space spanned by the rows of  $L_t$ . If  $F_t' \in R_t$ , then we can have an estimate for  $y_t$ , and we can, upon observing the value of  $y_t$ , obtain the posterior distributions for  $L_t\theta_t$  and  $\phi$ . If, on the other hand,  $F_t' \notin R_t$ , then,  $y_t$ , which is linked to  $F_t'$  through (3.8.1), will bring information about a subspace which is not included in  $R_t$ . Using this information, we will enlarge  $R_t$ , such that,  $R_t$  will be a proper subspace of  $R_{t+1}$ . This will be handled by the inclusion of an additional row in  $L_t$ , linearly independent of the previous rows. The stopping criterion is, of course,  $r_t = n$ , and, once this is reached, we can simply turn to the standard updating equations.

Also, in our approach, we treat the case where  $G_t$  is a singular matrix with a particular form, as opposed to Pole and West approach, which has to suppose non-singularity of the system matrix. We shall, then, divide our study into two cases, depending on invertibility of the system matrix.

CASE 1:  $G_t$  is non-singular, for all  $t$ .

We first suppose  $F'_t \in R_t$ . Let  $L_t^C$  be any conditional inverse for  $L_t$ . For example,  $L_t^C$  may be given by  $L'_t(L_t L'_t)^{-1}$ , since  $L_t$  is a full row rank matrix. Then, we have:

$$L'_t L_t^{C'} F_t = F_t,$$

which can also be written as:

$$F'_t(I - L_t^C L_t) = 0$$

Moreover, there is a unique vector  $\ell_t \in \mathbf{R}^{r_t}$  such that  $F'_t = \ell'_t L_t$ , given by:

$$\ell_t = L_t^{C'} F_t$$

Hence, (3.8.1) may be rewritten as:

$$y_t = \ell'_t L_t \theta_t + v_t$$

and we can have an estimate  $\hat{y}_t$  of  $y_t$ , in this case.

Let  $\theta_t^* = L_t \theta_t$ . At time  $(t - 1)$  we have the prior information:

$$\begin{aligned} (y_t | \phi, \theta_t^*, D_{t-1}) &\sim N[\ell'_t \theta_t^*; \phi^{-1}] \\ (\ell'_t \theta_t^* | \phi, D_{t-1}) &\sim N[\ell'_t a_t; \ell'_t P_t^{-1} \ell_t / \phi] \\ (\phi | D_{t-1}) &\sim G(n_{t-1}/2; n_{t-1} S_{t-1}/2) \end{aligned}$$

Hence, if we let:

$$e_t = y_t - \ell'_t a_t \tag{3.9}$$

$$Q_t^* = 1 + \ell'_t P_t^{-1} \ell_t, \tag{3.10}$$

we have the parameters of the posterior distribution  $G(n_t/2; n_t S_t/2)$  for  $(\phi | D_t)$ , being updated by:

$$\begin{aligned} n_t &= n_{t-1} + 1 \\ n_t S_t &= n_{t-1} S_{t-1} + e_t^2 / Q_t^* \end{aligned}$$

Also, conditional posterior distribution for  $\theta_t^*$  can be obtained from the prior information:

$$\begin{aligned} (\theta_t^* | \phi, D_{t-1}) &\sim N[a_t; (\phi P_t)^{-1}] \\ (y_t | \phi, D_{t-1}) &\sim N[\ell'_t a_t; Q_t^* / \phi] \end{aligned}$$

and:

$$\text{Cov}(y_t, \theta_t^* | \phi, D_{t-1}) = \text{Cov}(\ell_t' \theta_t^* + v_t, \theta_t^* | \phi, D_{t-1}) = \ell_t' P_t^{-1} / \phi$$

Then,  $(\theta_t^* | \phi, D_t) \sim N[a_t^*; (\phi P_t^*)^{-1}]$ , where:

$$a_t^* = a_t + P_t^{-1} \ell_t e_t / Q_t^*$$

$$P_t^* = P_t + \ell_t \ell_t'$$

and  $e_t, Q_t^*$ , are given by (3.9) and (3.10). Thus, we have the posterior distribution for  $\theta_t^*$ .

Now, we calculate the prior distribution for the next step. By (3.8.2):

$$G_t^{-1} \theta_{t+1} = \theta_t + G_t^{-1} \omega_{t+1}, \quad (3.11)$$

which gives:

$$L_t G_t^{-1} \theta_{t+1} = \theta_t^* + L_t G_t^{-1} \omega_{t+1} \quad (3.12)$$

Define:

$$L_{t+1} = L_t G_t^{-1} \quad (3.13)$$

Then, by (3.12) and (3.13),  $\theta_{t+1}^* = L_{t+1} \theta_{t+1}$  may be rewritten as:

$$\theta_{t+1}^* = \theta_t^* + L_{t+1} \omega_{t+1}$$

Thus, our prior distribution is  $(\theta_{t+1}^* | \phi, D_t) \sim N[a_{t+1}^*; (\phi P_{t+1})^{-1}]$ , with:

$$a_{t+1}^* = a_t^*$$

$$P_{t+1}^{-1} = P_t^{*-1} + L_{t+1} W_{t+1}^* L_{t+1}'$$

Now, we suppose  $F_t' \notin R_t$ . Then,  $y_t$  will bring information about a larger subspace, and will enlarge the matrix  $L_t$ .

Pre-multiplying  $F_t'$  in (3.11) and combining the result with (3.12), gives:

$$\begin{pmatrix} L_t \\ F_t' \end{pmatrix} G_t^{-1} \theta_{t+1} = \begin{pmatrix} L_t \\ F_t' \end{pmatrix} \theta_t + \begin{pmatrix} L_t \\ F_t' \end{pmatrix} G_t^{-1} \omega_{t+1}$$

Now, we define:

$$L_{t+1} = \begin{pmatrix} L_t \\ F_t' \end{pmatrix} G_t^{-1}$$

which gives:

$$L_{t+1} \theta_{t+1} = \begin{pmatrix} L_t \\ F_t' \end{pmatrix} \theta_t + L_{t+1} \omega_{t+1} \quad (3.14)$$

Now, from (31), we obtain  $a_{t+1}$  and  $P_{t+1}^{-1}$ , as:

$$a_{t+1} = \begin{pmatrix} a_t \\ y_t \end{pmatrix} \quad (3.15)$$

$$P_{t+1}^{-1} = \begin{bmatrix} P_t^{-1} & 0 \\ 0 & 1 \end{bmatrix} + L_{t+1}W_{t+1}L_{t+1}' \quad (3.16)$$

$F_t'$  is linearly independent of the rows of  $L_t$ ; therefore,  $F_t'\theta_t$  has an infinite variance. Also, we know that the covariance between  $L_t\theta_t$  and  $F_t'\theta_t$  is finite. Beside,  $L_t\theta_t$  and  $y_t$  are conditionally independent, given  $F_t'\theta_t$ , since the error in equation (3.8.1) is uncorrelated with  $\theta_t$ . Therefore, because of the invariance of the correlational structure with the conditioning, we can conclude that  $F_t'\theta_t$  and  $L_t\theta_t$  will be uncorrelated, given the value of observation  $y_t$ . This justifies the diagonal form in (3.16).

CASE 2:  $G_t$  is a singular matrix.

For this case, we will suppose that  $G_t$  assumes the form:

$$G_t = \begin{bmatrix} H_t & 0 \\ 0 & J_k(0) \end{bmatrix}$$

where  $J_k(0)$  is a  $k \times k$  Jordan block corresponding to a zero eigenvalue,  $k > 1$ , and  $H_t$  is nonsingular. We write commensurably:

$$\begin{aligned} L_t &= \begin{bmatrix} U_t & 0 \\ 0 & I_k \end{bmatrix} \\ \theta_t' &= [\psi_t' \quad \eta_t']' \\ F_t' &= [X_t' \quad e_{k,1}']' \end{aligned}$$

where  $I_k$  is the  $k \times k$  identity matrix and  $e_{k,1}$  corresponds to the first vector in the canonical basis of  $\mathbf{R}^k$ .

If  $M$  is a matrix with  $n$  rows (columns), we will denote by  $M^a(M^d)$ , a matrix obtained from  $M$ , by eliminating the  $(n+1-k)$ -th row (column).

Let  $J_k'(0)$  denote the transpose of  $J_k(0)$ .  $J_k'(0)$  is the generalized (or Moore Penrose) inverse of  $J_k(0)$  (see, e.g., Rao, 1962). Therefore, the Moore Penrose inverse,  $G_t^-$ , of  $G_t$ , is:

$$G_t^- = \begin{bmatrix} H_t^{-1} & 0 \\ 0 & J_k'(0) \end{bmatrix} \quad (3.17)$$

From (3.8.2) we get:

$$G_t^-\theta_{t+1} = G_t^-G_t\theta_t + G_t^-\omega_{t+1}$$

As the  $(n + 1 - k)$ -th row of  $G^-$  is a row of zeros, the  $(n + 1 - k)$ -th equation above will trivially be given as  $0 = 0$ . Then we drop this equation, to obtain:

$$(G_t^-)^a \theta_{t+1} = (G_t^- G_t)^a \theta_t + (G_t^-)^a \omega_{t+1}$$

But  $G_t^- G_t$  differs from the identity matrix, only by the  $(n + 1 - k)$ -th element in the diagonal, which is equal to zero. Hence, we can write:

$$(G_t^-)^a \theta_{t+1} = \theta_t^a + (G_t^-)^a \omega_{t+1}$$

Because of (3.8.2) and the form of  $G_t$ , the last component in  $\theta_{t+1}$  is the same as in  $\omega_{t+1}$ ; therefore:

$$\begin{bmatrix} (G_t^-)^a \\ e'_n \end{bmatrix} \theta_{t+1} = \begin{pmatrix} \theta_t^* \\ 0 \end{pmatrix} + \begin{bmatrix} (G_t^-)^a \\ e'_n \end{bmatrix} \omega_{t+1} \quad (3.18)$$

If  $F'_t \in R_t$ , we obtain the posterior  $(\theta_t^* | \phi, D_t)$  as usual (see the anterior case).

If this is not the case, then we can write:

$$\left( \begin{pmatrix} L_t \\ F'_t \end{pmatrix} \theta_t \middle| \phi, D_t \right) \sim N \left\{ \begin{pmatrix} a_t \\ y_t \end{pmatrix}; \begin{bmatrix} P_t^{-1} & 0 \\ 0 & 1 \end{bmatrix} \phi^{-1} \right\} \quad (3.19)$$

But, observe that:

$$\begin{pmatrix} L_t \\ F'_t \end{pmatrix} \theta_t = \begin{pmatrix} U_t \psi_t \\ \eta_t \\ X'_t \psi_t + \eta_{1,t} \end{pmatrix}$$

Consider the matrix  $M_t$ , defined as:

$$M_t = \begin{bmatrix} \begin{pmatrix} U_t \\ X'_t \end{pmatrix} & 0 \\ 0 & I_k \end{bmatrix}$$

We have:

$$M_t \theta_t = \begin{pmatrix} U_t \psi_t \\ X'_t \psi_t \\ \eta_t \end{pmatrix}$$

From (3.19), we can easily obtain  $(M_t \theta_t | \phi, D_t)$ . In fact, from the diagonal structure we derive:

$$Cov(U_t \psi_t, X'_t \psi_t | \phi, D_t) = -Cov(U_t \psi_t, \eta_{1,t} | \phi, D_t)$$

$$Var(X'_t \psi_t | \phi, D_t) = \phi^{-1} + Var(\eta_{1,t} | \phi, D_t)$$

$$Cov(X'_t \psi_t, \eta_t | \phi, D_t) = -Cov(\eta_{1,t}, \eta_t | \phi, D_t)$$

Define  $Z_t$  as  $L_t$  or  $M_t$ , accordingly to the first or second case. We have the conditional distribution for  $(Z_t \theta_t | \phi, D_t)$ . Let

$$Z_t = \begin{bmatrix} B_t & 0 \\ 0 & I_k \end{bmatrix} \quad (3.20)$$

where  $B_t$  will be  $U_t$  or  $[U_t' \ X_t']'$ , accordingly.

Consider the matrix:

$$\begin{bmatrix} Z_t^d & 0 \\ 0 & 1 \end{bmatrix} \quad (3.21)$$

Pre-multiplying (3.21) in (3.18), we get:

$$\begin{bmatrix} Z_t^d(G_t^-)^a \\ e_n' \end{bmatrix} \theta_{t+1} = \begin{pmatrix} Z_t^d \theta_t^a \\ 0 \end{pmatrix} + \begin{bmatrix} Z_t^d(G_t^-)^a \\ e_n' \end{bmatrix} \omega_{t+1} \quad (3.22)$$

From (3.17) and (3.20), we can conclude:

$$Z_t^d(G_t^-)^a = \begin{bmatrix} B_t H_t^{-1} & 0 \\ 0 & I_k^d \{J_k'(0)\}^a \end{bmatrix} = \begin{bmatrix} B_t H_t^{-1} & 0 \\ 0 & J_k'(0) \end{bmatrix}$$

Defining

$$K_{t+1} = \begin{bmatrix} Z_t^d(G_t^-)^a \\ e_n' \end{bmatrix} \quad (3.23)$$

we have, by (3.22) and (3.23):

$$K_{t+1} \theta_{t+1} = \begin{pmatrix} Z_t^d \theta_t^a \\ 0 \end{pmatrix} + K_{t+1} \omega_{t+1} \quad (3.24)$$

Observing that

$$K_{t+1} = \begin{bmatrix} B_t H_t^{-1} & 0 \\ 0 & 0 \\ 0 & I_k \end{bmatrix} \quad Z_t^d \theta_t^a = \begin{pmatrix} B_t \psi_t \\ 0 \\ \eta_t^a \end{pmatrix} \quad (3.25)$$

we see, by (3.25) that the  $(n+1-k)$ -th equation in (3.24) is trivially  $0 = 0$ . Then, we may drop this equation, defining:

$$L_{t+1} = K_{t+1}^a = \begin{bmatrix} B_t H_t^{-1} & 0 \\ 0 & I_k \end{bmatrix}$$

with the same forme as  $L_t$ .

Dropping this equation, we will get:

$$L_{t+1} \theta_{t+1} = \begin{pmatrix} B_t \psi_t \\ \eta_t^a \\ 0 \end{pmatrix} + L_{t+1} \omega_{t+1}$$

and since we know the distribution of  $(Z_t \theta_t | \phi, D_t)$ , we may calculate that one for  $(L_{t+1} \theta_{t+1} | \phi, D_t)$ , and go to the next step.

The procedure we propose here uses, therefore, the first observations much more efficiently, since it considers the initial information that can be provided by the model, in order to obtain an initial proper prior distribution, at least for a subspace

of the state vector. This subspace can be enlarged, as we collect more information along time. At a certain instant of time we have a proper distribution for  $\theta_t^* = L_t \theta_t$ . Then, according to  $F_t$ , we can enlarge the subspace for which we have a proper distribution (when  $F_t \notin R_t$ ), or we can use the observation  $y_t$  to obtain a proper posterior distribution in that particular subspace with a bigger precision (when  $F_t \in R_t$ ). This gives a more efficient use of the initial information provided by the model and observations. We must note that if the matrix  $L_t$  is never enlarged, which means  $F_t \in L_t$ , this is of no forecasting importance, as long as we can obtain proper conditional distributions for the future observations (if, for example,  $F_{t+1}, F_{t+2}, \dots, F_{t+k}$  belong to the space spanned by the columns of  $L_t'$ , then, we can find vectors  $\ell_{t+1}, \ell_{t+2}, \dots, \ell_{t+k}$ , such that  $F_{t+i} = L_{t+i}' \ell_{t+i}, i = 1, \dots, k$ ; then, the forecast distributions can be obtained from the information on  $\theta_t^*$ ) and this is our prime objective. For this case, the precision of the distribution of  $\theta_t^* = L_t \theta_t$  is increased, as we collect information. Of course, in control situations it may well be required that a proper prior for the full space be quickly obtained. This is one of the fundamental points behind the development in Box and Draper (1969).



## CHAPTER 4

### THE BOOKING MODEL

#### 4.1 Introduction

In this chapter we propose a model to forecast the distribution of the number of passengers booking tickets for particular flights. As we have mentioned in the first chapter, the most commonly used models for these purposes present some drawbacks, the most serious one being the fact that normal distribution is widely assumed for data. This does not provide a good fit, since we are often dealing with *small discrete values*. Another important feature is that it must be a dynamic model, in the sense that it must be corrected for special events. In this chapter we define the proposed model and develop a method for updating the forecast distributions of the booking numbers. As part of the method, we construct a log-normal to gamma approximation, minimizing the  $L_2$  distance between these two densities. We present the updating procedure, and also some simulated results which give an indication that the density approximation we have developed is well applied to our specific problem. The aggregation problem, which will be discussed in the next chapter is also mentioned. Finally, in the appendix, we present the details of the gamma to log-normal approximation used.

#### 4.2 The booking problem

We begin by constructing a model to handle discrete data assuming relatively small values. In this context, the Poisson structure looks to be convenient for our purposes. We will suppose that data, given a mean parameter  $\lambda$ , must follow a Poisson distribution with mean proportional to  $\lambda$ .

It looks sensible to consider a different model for each particular regular flight, to begin with. Let's take a specific flight (e.g., from London to New York), and suppose that this service obeys a regular routine. In other words, there exists a fixed and constant interval of time between the beginning of bookings and date of departure, for this flight. Also, the flight takes off on a regular basis (e.g., every Thursday), such that the interval of time between two consecutive flights is constant. A forecast for the number  $x$  of passengers reserving seats in the plane is required, so that the company can take decisions with respect to maximizing expected income from that specific flight. The information about  $x$  can be updated over time, through a Bayesian approach. Let's divide the booking period into  $k$  different blocks, and we take the random variables  $X_i, i = 1, \dots, k$  of reservations for each block as

$$(X_i|\varphi) \sim Po[r_i\varphi] \tag{4.1}$$

where the  $r_i$ 's define the *reservation curve*;  $r_i$  is the proportion of people booking tickets in the  $i$ -th period of time. Normally, the time interval between two consecutive flights is one week, hence, for simplicity of exposition, consider the booking period divided into weeks, and refer to the period of time of one observation as one week. We have  $0 < r_i < 1, \forall i$  and  $\sum r_i = 1$ . Here,  $\varphi$  is a 'demand level parameter' for which we assume an initial gamma prior distribution with parameters  $\alpha$  and  $\beta$ , the probability density of  $\varphi \sim G(\alpha, \beta)$  being given by:

$$f(\varphi|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \varphi^{\alpha-1} e^{-\beta\varphi} \quad (4.2)$$

The prior distribution for  $\varphi$ , as well as the reservation curve, will be based on similar flights, i.e., using past experience with that particular flight. In our model, we consider the reservation curve to be the same for all flights, i.e., the probability distribution along the booking period does not change from flight to flight. This looks to be a reasonable assumption, at least locally, if the time between two consecutive flights is relatively small, as is the case with weekly flights.

Observe that data represents the total number of people booking seats in particular flights. It is important to note that this data may be used to infer about the total number  $\tilde{x}$  of people confirming seats. In fact, let  $h_i$  be the probability that people who book seats will really confirm the reservation. Then, we simply have:

$$(\tilde{x}_i|\varphi) \sim Po(h_i r_i \varphi)$$

that means we can use the methodology just described by defining  $\tilde{r}_i = r_i h_i$ .

Another important feature to be incorporated in the model is the controls to which data is submitted. For example, the company may decide, for some special reasons, to close a particular class, or a certain terminal may be temporarily closed if it is less profitable than others and bookings are high. Therefore, if a class has to be closed, we can consider a parameter  $\gamma_i$  representing the proportion of seats in the plane corresponding to the open classes (that means, the total number of seats corresponding to the open classes with respect to the total number of seats in the plane). Also, some terminals may be closed, and we can think of  $\delta_i$  as being the expected fraction of booking from the open terminals. In this case, the number  $x_i^*$  of people really taking the flight is distributed as:

$$(x_i^*|\varphi) \sim Po(r_i^* \varphi)$$

where  $r_i^* = \tilde{r}_i \gamma_i \delta_i$ .

Let's consider the variable  $x_i$  defined in (4.1). It is well known that, given (4.1) and (4.2), the marginal distribution for  $x_i$  will be a negative binomial:

$$x_i \sim NB(\alpha, p) \quad (4.3)$$

where  $p = \beta/(\beta + r_i)$ . Observe that  $p$  is a crescent function of  $\beta$ , something that looks sensible if we consider that the distribution above is the distribution of the number of failures after  $\alpha$  successes have been obtained in a sequence of independent trials, each with probability  $p$  of success. Then, for fixed  $\alpha$ , a larger probability of success must correspond to a smaller expected number of trials until we get exactly  $\alpha$  successes, and, consequently, a smaller expected value of the negative binomial. This is really true; if we observe that

$$\mu_i = E[x_i] = E\{E[x_i|\varphi]\} = r_i E[\varphi] = r_i \alpha / \beta$$

we can see that  $\mu_i$  is a decreasing function of  $\beta$ , and, therefore, a decreasing function of  $p$ .

The updated (posterior) distribution of  $\varphi$ , after observing  $x_1, \dots, x_k$ , is, according to well known conjugate analysis results, given by:

$$(\varphi|x_1, \dots, x_k) \sim G(\alpha^*, \beta^*)$$

where  $\alpha^* = \alpha + S_k$  and  $\beta^* = \beta + R_k$ ; the quantities  $S_k$  and  $R_k$  being given by the partial sums  $S_k = \sum_{i=1}^k x_i$  and  $R_k = \sum_{i=1}^k r_i$ . From these equations we can readily see that the posterior mean  $\mu^*$  can be written as:

$$\mu^* = p\mu_0 + q\hat{\mu} \quad (4.4)$$

where  $\mu_0$  correspond to the prior mean,  $\hat{\mu} = S_k/R_k$  is a natural estimator of the mean of  $\varphi$  based upon the sample,  $p = \beta/(\beta + R_k)$  is the weight given to the prior mean  $\mu_0$  and  $p + q = 1$ . By a simple argument which is similar to the one we have presented before, we can see that  $p$  is reasonably a crescent function of  $\beta$ . Also, the largest the sample size  $k$ , the largest the value of  $R_k$ , therefore,  $p$  must be a decreasing function of  $R_k$ . We also observe that (4.4) can be put in the form:

$$\mu^* = \mu_0 + q(\hat{\mu} - \mu_0)$$

meaning that the posterior mean is the prior mean added by a 'bias' corrected by a factor which gets bigger as the sample becomes more informative.

Let  $D_t$  represent all the relevant information up to time  $t$ . Then, we forecast the number of reservations at time  $t + 1$  by:

$$\hat{x}_{t+1} = E[x_{t+1}|D_t] = E\{E[x_{t+1}|\varphi]|D_t\} = E[r_{t+1}\varphi|D_t] = r_{t+1}\mu_{t+1} \quad (4.5)$$

where  $\mu_{t+1}$  is the mean of posterior distribution of  $\varphi$  after  $t$  updatings.

Now, if we write  $(\varphi|D_t) \sim G(\alpha_t, \beta_t)$ , then (4.5) can be rewritten as:

$$r_t \hat{x}_{t+1} = r_{t+1} \{p_t \hat{x}_t + q_t x_t\}$$

where  $p_t + q_t = 1$  and  $q_t = \beta_{t-1}/\beta_t$ . Again, if  $q_t$  is very small, we should expect, in principle, that the variance for the distribution of  $\varphi$  has a significant decrease after the updating, meaning that we should give a reasonable weight to the estimate  $\hat{x}_t$ . Observe that the sequence  $\{p_t\}$  (and therefore  $\{q_t\}$ ) does not depend on the observed values and can therefore be known *a priori* if we have the reservation curve  $\{r_t\}$ .

Now, the variances  $V_t = Var[x_{t+1}|D_t]$  can be easily obtained as:

$$\begin{aligned} V_t &= V[x_{t+1}|D_t] = E[V\{x_{t+1}|\varphi\}|D_t] + V[E\{x_{t+1}|\varphi\}|D_t] \\ &= E[r_{t+1}\varphi|D_t] + V[r_{t+1}\varphi|D_t] \\ &= \hat{x}_t + r_{t+1}\hat{x}_t/\beta_t = \hat{x}_t/q_{t+1} \end{aligned}$$

This gives an idea of how we can update our distributions by use of a conjugate analysis.

We shall make the above model more sophisticated in order to consider the time variation of  $\varphi$ , in such a way that there is a correlation structure linking flights departing on consecutive weeks. In this case, the information about bookings of a specific flight will be used to update the prior beliefs about other flights; modification of beliefs about  $\varphi_t$  (posterior distribution) implies modification of beliefs about  $\varphi_{t+1}$ , provided there is a known and fixed correlation structure linking these two quantities.

### 4.3 Model definition

We shall let  $x_{it}$  be the number of occurrences during the  $i$ -th period for the  $t$ -th process. In our specific example, it will be the number of booked seats during the  $i - th$  week of the booking period referring to the flight departing at time  $t$ . We consider a real parameter  $\varphi_t$ , such that, conditional on  $\varphi_t$ , the variables  $x_{1t}, \dots, x_{kt}$  are independently and Poisson distributed. We will also suppose that:

$$E[x_{it}|\varphi_t] = r_i \varphi_t$$

where we will put  $\varphi_t \sim G(\alpha_t, \beta_t)$  in order to have conjugacy and where  $\{r_1, \dots, r_k\}$ , which will be called the reservation curve, is a set of known constants, independent of  $t$ , such that  $r_i > 0, \forall i$  and  $\sum r_i = 1$ . The reservation curve defines, therefore, the relative proportion of occurrences for the  $i$ -th period.

Using the notation of Chapter 2 in the discussion of the DGLM we have the following set up. We define a state space vector  $\theta_t$ , evolving in time, according to the system equation:

$$\theta_t = \theta_{t-1} + \omega_t \quad (4.6)$$

where the error sequence  $\omega_t$  is independent. We assume  $\omega_t$  is  $N[0, W_t]$  distributed, where  $W_t$  is known for all  $t$ . We will use in our model the discount factor approach, where we suppose that  $Var[\theta_t | D_{t-1}] = Var[\theta_{t-1} | D_{t-1}] / \delta$ , where  $\delta$  is the fixed known discount factor.

We shall consider, as well, a vector sequence of regressors,  $F_t$ , such that:

$$\lambda_t = F_t' \theta_t \quad (4.7)$$

We introduce here  $\eta_t = \log \varphi_t$ , the natural parameter for the Poisson distribution of the total number of occurrences in the  $t$ -th process. The univariate parameter  $\lambda_t$  will be linked with  $\eta_t$  through a known bijection  $f(\cdot)$ , between gamma and log-normal distributions, such that  $\lambda_t = f(\eta_t)$ .

At this point, we have our model in the Dynamic Generalized Linear Model framework. Equation (4.7) constitutes a regression on the log-scale, in the sense that we try to explain the log-level (log-total number of occurrences for the  $t$ -th process) by the components of the vector  $F_t$ , which is the vector of explanatory variables. In our problem,  $F_t$  will be a vector of variables that influences the total number of booked seats, such as price of the airline ticket, cost of the optional ways of transport, and so on.

We could think of using a simpler model where  $\lambda_t$  is directly identified with  $\eta_t$ ; in other words, we could think of a model where  $f(\cdot)$  is simply the identity mapping. This would simply fit, then, a Poisson-lognormal distribution for the observations. This distribution has been introduced in Preston (1948) and has been further studied, by Bulmer (1974), Reid (1981), Aitchison and Ho (1989), among others. However, it is not good for our purposes, since we are interested here in a closed sequential updating, and this means we must have conjugacy.

#### 4.4 Simple updating

Very briefly, we can consider the updating procedure for this model, given an observation, in the following way:

1) We are given the posterior distribution for the state vector  $\theta_{t-1}$ . Let  $D_{t-1}$  represent all the relevant information available at time  $t - 1$ . Then, we have:

$$(\theta_{t-1}|D_{t-1}) \sim N[m_{t-1}, C_{t-1}]$$

where the quantities  $m_{t-1}$  and  $C_{t-1}$  are known.

2) We consider a discount factor  $\delta$ , such that the prior distribution for  $\theta_t$  is given by:

$$(\theta_t|D_{t-1}) \sim N[a_t, R_t] \quad (4.8)$$

with  $a_t = m_{t-1}$  and  $R_t = C_{t-1}/\delta$ .

3) Now, according to (4.7), the prior distribution for  $\lambda_t$ , conditional on  $D_{t-1}$ , will be given as:

$$(\lambda_t|D_{t-1}) \sim N[f_t, q_t] \quad (4.9)$$

with  $f_t = F_t' a_t$  and  $q_t = F_t' R_t F_t$ .

4) The bijection is defined so that we can obtain the prior distribution for  $\varphi_t$ , from the distribution above for  $\lambda_t$  as

$$(\varphi_t|D_{t-1}) \sim G[\alpha_t, \beta_t]$$

where  $\alpha_t$  and  $\beta_t$  are functions of  $f_t$  and  $q_t$ .

5) Now, observe  $x_{it}$ , the number of occurrences for the  $i$ -th period of the  $t$ -th process. By simple conjugate analysis, we update parameters of the gamma distribution as:

$$(\varphi_t|D_t) \sim G[\alpha_t^*, \beta_t^*] \quad (4.10)$$

where  $\alpha_t^* = \alpha_t + x_{it}$  and  $\beta_t^* = \beta_t + r_i$ .

6) Now, obtain posterior distribution for  $\lambda_t$  via bijection:

$$(\lambda_t|D_t) \sim N[f_t^*, q_t^*]$$

where  $f_t^*$  and  $q_t^*$  are functions of  $\alpha_t^*$  and  $\beta_t^*$ .

7) Posterior distribution for time  $t$  can now be obtained:

$$(\theta_t|D_t) \sim N[m_t, C_t]$$

where  $m_t$  and  $C_t$  are obtained by:

$$\begin{aligned} m_t &= E[\theta_t|D_t] = E[E\{\theta_t|\lambda_t, D_{t-1}\}|D_t] \\ &= E[a_t + R_t F_t(\lambda_t - f_t)/q_t|D_t] \\ &= a_t + R_t F_t(f_t^* - f_t)/q_t \end{aligned} \quad (4.11.1)$$

and

$$\begin{aligned}
C_t &= V[\theta_t|D_t] = V[E\{\theta_t|\lambda_t, D_{t-1}\}|D_t] + E[V\{\theta_t|\lambda_t, D_{t-1}\}|D_t] \\
&= V[a_t + R_t F_t(\lambda_t - f_t)/q_t|D_t] + E[R_t - R_t F_t F_t' R_t/q_t|D_t] \\
&= R_t F_t F_t' R_t q_t^*/q_t^2 + R_t - R_t F_t F_t' R_t/q_t \\
&= R_t - R_t F_t F_t' R_t(1 - q_t^*/q_t)/q_t
\end{aligned} \tag{4.11.2}$$

Defining  $A_t = R_t F_t/q_t$ , we can write:

$$m_t = a_t + A_t(f_t^* - f_t) \tag{4.12.1}$$

$$C_t = R_t - A_t A_t'(q_t - q_t^*) \tag{4.12.2}$$

The new information enters the updating equations above via  $f_t^*$  and  $q_t^*$ .

The evolution of the parameters  $m_t$  and  $C_t$  depends on a crucial way of the form of evolution of  $f_t^*$  and  $q_t^*$ , since these are the only quantities in the equations (4.12.1) and (4.12.2) depending on the new observation  $x_{it}$ . From this fact, we can see the fundamental importance of the bijection linking the distributions of  $\eta_t$  and  $\lambda_t$ ; different relationships between the distributions can lead to different values of  $f_t^*$  and  $q_t^*$ .

It is important, at this point, to stress that, although we refer to the bijection between  $\lambda_t$  and  $\eta_t$  as a known function  $f(\cdot)$  which remains the same during all the updating procedure, we are, in fact, using an approximation. The gamma distribution for  $\varphi_t$  and the log-normal distribution for  $\lambda_t$  are being linked, in our procedure, through their parameters. Therefore, it must be clear that the function linking these two parameters will possibly depend (and will depend) on these parameters, and does not remain the same in each updating step. We can, however, expect, that, as the updating progresses, the approximation method will become more precise, in the sense that there will be a small difference between these functions, and the method will, then, become approximately coherent (for coherence of the DGLM, see Smith (1992)).

Step 6 of the updating must, therefore, be seen, in practice, as an approximation, as Step 7 works as the best estimation of the first two moments of the posterior distribution for  $\theta_t$ .

To obtain the forecasting equations, we consider the prior distribution of  $\lambda_{t+k}$ , given the information up to time  $t$ :

$$(\lambda_{t+k}|D_t) \sim N[f_t(k), q_t(k)]$$

where the parameters above are given by:

$$\begin{aligned} f_t(k) &= F'_{t+k} m_t \\ q_t(k) &= F'_{t+k} \left( C_t + \sum_{i=1}^k W_{t+i} \right) F_{t+k} \end{aligned}$$

From the distribution above and the bijection, we, again, use the approximation:

$$(\varphi_{t+k}|D_t) \sim G(\alpha_t(k), \beta_t(k))$$

where  $\alpha_t(k)$  and  $\beta_t(k)$  are functions of  $f_t(k)$  and  $q_t(k)$ .

Now, we forecast  $x_{i,t+k}$  by:

$$\begin{aligned} x_{i,t}(k) &= E[x_{i,t+k}|D_t] = E[E\{x_{i,t+k}|\varphi_{t+k}\}|D_t] \\ &= r_i E[\varphi_{t+k}|D_t] = r_i \alpha_t(k) / \beta_t(k) \end{aligned}$$

Also, the prior variance of the distribution can be obtained by:

$$\begin{aligned} V_{i,t}(k) &= V[x_{i,t+k}|D_t] \\ &= V[E\{x_{i,t+k}|\varphi_{t+k}\}|D_t] + E[V\{x_{i,t+k}|\varphi_{t+k}\}|D_t] \\ &= r_i^2 V[\varphi_{t+k}|D_t] + r_i E[\varphi_{t+k}|D_t] \\ &= r_i \alpha_t(k) (1 + r_i / \beta_t(k)) / \beta_t(k) \\ &= (1 + r_i / \beta_t(k)) x_{i,t}(k) \end{aligned}$$

#### 4.5 The choice of $f(\cdot)$

The link function relating  $\lambda_t$  to  $\eta_t$  plays a fundamental role in our model, since it determines the form by which  $f_t^*$  and  $q_t^*$  will be obtained. In our model,  $\lambda_t$  is an “approximation” for  $\eta_t$ , in the sense that  $\lambda_t$  is the regressed log-total number of outcomes. Hence, it is appropriate to consider a bijection between these two parameters, such that their distributions are quite near each other, in some special sense. With this idea in mind, we can try, for example, to approximate the density of a gamma distribution by a log-normal density or vice-versa.

A first idea is to equate mean and variance for both distributions. We shall expect this method to work very efficiently for a relatively small coefficient of variation. In fact, when this is the case, both distributions can be very well approximated by a normal distribution, and both curves will fall very near the normal curve with that same mean and variance. On the other hand, we shall expect both curves to be completely different for a large value of the coefficient of variation. For example, the



gamma density is not a bounded function, if the coefficient of variation is greater than unit, while the log-normal density is always a bounded function.

Because the idea above does not produce very good results when the coefficient of variation is relatively large (for a coefficient of variation of 0.5, curves show a reasonable difference between each other), we must try another method to construct the desired relationship between distributions. The method we have adopted is the numerical minimization of the  $L_2$  distance between the densities. Then, if  $p(\cdot)$  is a gamma density with coefficient of variation  $k_G$ , such that  $k_G^2 < 2$ , we try to minimize the integral of the squares of residuals, given by

$$d_{L_2}(p, g) = \int_0^\infty |p(x) - g(x)|^2 dx \quad (4.13)$$

for all log-normal densities  $g(\cdot)$ . It can be easily seen that the integral above is finite if and only if  $k_G^2 < 2$ .

We have implemented this idea, by use of a numerical approach to minimize the integral. In our study, we have worked with an original gamma distribution of unit mean, and considered the approximation for different values of  $k_G$ . The conclusion was that the log-normal density obtained by numerical minimization of (4.13) can be used as a good approximation for the original gamma density, if  $k_G \leq 0.5$ . Observing the results for twelve different values of  $k_G$ , we tried to obtain an analytical relationship, giving approximately the mean and coefficient of variation of the best log-normal obtained, for each gamma density we fix. The two fitted curves are given by:

$$\mu_L = 1 + 0.2886k_G^2 \quad (4.14.1)$$

$$k_L = 0.9135k_G + 0.4477k_G^2 \quad (4.14.2)$$

where  $\mu_L$  and  $k_L$  are, respectively, the mean and coefficient of variation of the best log-normal obtained. Equation (4.14.1) may be extended to

$$\mu_L = \mu_G(1 + 0.2886k_G^2) \quad (4.15)$$

where  $\mu_G$  is the mean for a general gamma density.

The results obtained by use of (4.14.2) and (4.15) above were observed to be as good as those obtained by direct minimization. Therefore, the bijection we derived can be considered a good solution to our approximation problem. It is important to stress that the significant difference between the densities for a reasonably large value of  $k_G$ , does not constitute a very serious problem. Indeed, a large value of

$k_G$  reflects a large uncertainty about the parameter and we do not lose a lot by not working with a very good approximation.

Details of the approximation idea above explained are given in the appendix to this chapter.

#### 4.6 Robustness of $f(\cdot)$

The parameter  $\lambda_t$  reflects our beliefs about the behaviour of  $\varphi_t$ , via the linear regression (4.7). It is important, then, that  $f(\cdot)$  must be robust in the sense that slight modifications in the distribution of  $\lambda_t$  will produce slight modifications in the distribution of  $\varphi_t$  and vice-versa. For example, we can see that the updating described in the anterior section is directly made in the distribution of  $\varphi_t$  (eq. (4.10)), while the time evolution is described by  $\lambda_t$ . Therefore, it is highly desirable that a small modification in the distribution of  $\varphi_t$ , which will occur if we have a strong gamma prior for this parameter, will yield a posterior distribution for  $\lambda_t$ , which is very near its initial prior distribution.

Let  $p$  be a  $LN(\mu_1, \sigma_1^2)$  density function and let  $g$  be a  $LN(\mu_2, \sigma_2^2)$  density function. Evaluating the integral given by (4.13), the expression obtained is that of a continuous function of  $\mu_1, \mu_2, \sigma_1$  and  $\sigma_2$ . Therefore, the  $L_2$  distance given by (4.13) is a continuous function of the parameters, in this case. That means that if the distance between the parameters of two different log-normal distributions is too small, then, the densities of these log-normal distributions must be reasonably near each other, in the sense that their  $L_2$  distance must be relatively small.

We conclude that the bijection between the gamma and log-normal distributions must be such that the log-normal parameters must be obtained from their associated gamma parameters by a continuous function with continuous inverse (this is the case, for example, of the specific bijection we have earlier mentioned). Then, a strong prior gamma distribution for  $\varphi_t$ , which implies a small change of the gamma parameters in the updating, will produce a small modification in the log-normal parameters, and, therefore, the prior and posterior distributions for the  $\lambda_t$  will be very near each other.

#### 4.7 The updating problem

An important feature of our model is that we have, at a given instant of time, new information concerning processes beginning at various instants of time. To be more explicit, let's suppose that  $k$ , as defined in Section 4.2, is equal to two. Then, we will have at time  $t$ , new information to update the distribution of  $\lambda_t$  and  $\lambda_{t-1}$ . The updated distribution clearly depends of the link we have defined between the

normal distribution of  $\lambda_t$  and the gamma distribution of  $\varphi_t$ .

A simple procedure would consist of updating the joint distribution of  $\lambda_t$  and  $\lambda_{t-1}$ , using one observation each time. Using this idea, we should proceed by the following steps:

- 1) We are given  $((\lambda_t, \lambda_{t-1})|D_{t-1})$
- 2) Observe, at time  $t$ ,  $x_{1,t}$  and  $x_{2,t-1}$
- 3) Update  $\lambda_t$  to obtain  $(\lambda_t|\{D_{t-1}, x_{1,t}\})$
- 4) Use the prior joint distribution to obtain  $((\lambda_t, \lambda_{t-1})|\{D_{t-1}, x_{1,t}\})$ ; the pre-updated distribution
- 5) With the distribution above, update to get  $(\lambda_{t-1}|D_t)$ , where  $D_t$  is given by  $D_t = \{D_{t-1}, x_{1,t}, x_{2,t-1}\}$
- 6) Using the covariance structure of pre-updated distribution and updated distribution above, obtain  $(\lambda_t|D_t)$
- 7) From the equation defining the evolution of  $\lambda_t$  we can obtain the joint distribution  $((\lambda_{t+1}, \lambda_t)|D_t)$

The procedure above is expected to work if the updated joint distribution does not depend on the order in which the components are updated. This is clearly the case, if components have zero correlation; then, information about the first component does not affect our beliefs about the second component and vice-versa. The order of updating is also expected to be irrelevant for a small value of the coefficient of variation of the log-normal. This must be true, since a small coefficient of variation for the log-normal implies a small variance for the underlying normal. Then, for this case, the prior information should dominate the information coming from the observations. That means, the posterior distributions obtained in each case will be approximately the same. However, we cannot say, *a priori*, that the updating is independent of the order we update the distributions, which means we have to consider a method to update the joint distribution at one time, using all the information available.

To perform the updating, we will use the concept of equivalent observation. Consider the following observational equation:

$$y_t = \lambda_t + \nu_t \quad (4.16)$$

where  $\nu_t$  is an independent error sequence, such that  $\nu_t \sim N[0, V_t]$ , for each  $t$ . From (4.9) and (4.16) we obtain:

$$(y_t|D_{t-1}) \sim N[f_t, q_t + V_t]$$

and, therefore:

$$(\lambda_t | y_t, D_{t-1}) \sim N[f_t^*, q_t^*]$$

where the parameters above will be given by:

$$f_t^* = f_t + q_t(y_t - f_t)/(q_t + V_t)$$

$$q_t^* = q_t V_t / (q_t + V_t)$$

Conversely,  $y_t$  and  $V_t$  may be obtained from  $f_t^*$  and  $q_t^*$ :

$$y_t = (q_t f_t^* - q_t^* f_t) / (q_t - q_t^*) \quad (4.17.1)$$

$$V_t = q_t^* q_t / (q_t - q_t^*) \quad (4.17.2)$$

The equivalent observation, together with its associated variance, is the one to yield the same posterior distribution that was obtained using the bijection. This device gives us a representation form in terms of the DLM to study the problem.

Now, suppose we want to update the multivariate distribution of  $\Lambda = (\lambda_1, \dots, \lambda_k)$ , where  $\lambda_1, \dots, \lambda_k$  are the  $k$  univariate parameters, about which we collect information. As we have seen, we can obtain, for each  $\lambda_i$ , an equivalent observation  $y_i$ , together with its associated variance  $v_i$ . The values of  $y_i$  and  $v_i$  will be obtained from the observations and the bijection linking  $\varphi_t$  and  $\lambda_t$ . We consider, then, a vector  $\mathbf{y} = (y_1, \dots, y_k)$  of equivalent observations, such that

$$\mathbf{y} = \Lambda + \mathbf{v}$$

where  $\mathbf{v}$  is a multivariate random error, distributed as  $N(0, V)$ ,  $V$  being the diagonal matrix of  $v_1, \dots, v_k$ . Then, if  $\Lambda \sim N[\mathbf{f}, Q]$ , we find:

$$E[\Lambda | \mathbf{y}] = \mathbf{f} + Q(Q + V)^{-1}(\mathbf{y} - \mathbf{f}) \quad (4.18.1)$$

$$V[\Lambda | \mathbf{y}] = Q - Q(Q + V)^{-1}Q \quad (4.18.2)$$

Now, if the prior distribution for  $\lambda_i$  is  $N[f_i, q_i]$  and its posterior distribution is  $N[f_i^*, q_i^*]$ , then we know from equations (4.17.1) and (4.17.2) that  $y_i$  and  $v_i$  are obtained as:

$$y_i = (q_i f_i^* - q_i^* f_i) / (q_i - q_i^*) \quad (4.19.1)$$

$$v_i = q_i^* q_i / (q_i - q_i^*) \quad (4.19.2)$$

Now, a problem will occur if we get a zero observation  $x_i$ . From (4.10) we can observe that we will not change the coefficient of variation of the gamma distribution

associated with  $\lambda_i$ , in this case. We also know that the coefficient of variation of a log-normal is given as a function only of the variance of its underlying normal distribution. Therefore, if the defined bijection relates the gamma and log-normal distributions, by relating their coefficients of variation, then, we will not change the variance of  $\lambda_i$  after the updating. That means we will be left with an infinite equivalent observation, associated with an infinite variance (since  $q_i^*$  will coincide with  $q_i$ ).

This problem can be overcome if we observe that  $(Q + V)^{-1}$ , as appears in equations (4.18.1) and (4.18.2), can be expanded as:

$$(Q + V)^{-1} = V^{-1} - V^{-1}(V^{-1} + Q^{-1})^{-1}V^{-1}$$

Now, from (4.19.1) and (4.19.2) we observe that, when  $q_i = q_i^*$ :

$$v_i^{-1}y_i = (f_i^* - f_i)/q_i^*$$

and this is finite, provided that  $q_i^* \neq 0$ . Then, our basic idea is to express updating in terms of the  $z_i = (f_i^* - f_i)/q_i^*$ , when  $y$  and  $V$  are both infinite, i.e., when  $x = 0$ .

For the general case, consider  $\Lambda' = (\Lambda'_1, \Lambda'_2)$ , where the index (1) refers to the components for which we have a zero observation  $x$ , the index (2) referring to the other components. We partition  $\mathbf{f}$ ,  $\mathbf{y}$ ,  $Q$  and  $V$ , accordingly as:

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix} \quad Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \quad V = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix}$$

Let

$$S = (Q + V)^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

Then, the entries above are obtained as:

$$S_{11} = (I + UQ_{12}ZQ_{21})U \tag{4.20.1}$$

$$S_{12} = -UQ_{12}Z \tag{4.20.2}$$

$$S_{21} = S'_{12} \tag{4.20.3}$$

$$S_{22} = Z \tag{4.20.4}$$

where

$$U = (Q_{11} + V_1)^{-1} = (I - V_1^{-1}(Q_{11}^{-1} + V_1^{-1})^{-1})V_1^{-1}$$

$$Z = (Q_{22} + V_2 - Q_{21}UQ_{12})^{-1}$$

Now, observe that if we let  $V_1^{-1} \rightarrow 0$  in equations (4.20.1) to (4.20.4) we will get:

$$\begin{aligned} S_{11}(\mathbf{y}_1 - \mathbf{f}_1) &= V_1^{-1} \mathbf{y}_1 = \mathbf{z}_1 \\ S_{12}(\mathbf{y}_2 - \mathbf{f}_2) &= 0 \\ S_{21}(\mathbf{y}_1 - \mathbf{f}_1) &= -(Q_{22} + V_2)^{-1} Q_{21} \mathbf{z}_1 \\ S_{22}(\mathbf{y}_2 - \mathbf{f}_2) &= (Q_{22} + V_2)^{-1} (\mathbf{y}_2 - \mathbf{f}_2) \end{aligned}$$

Considering also the asymptotic behaviour of the entries of  $S$  alone, we can see that equations (4.18.1) and (4.18.2) may be rewritten as:

$$E[\Lambda|\mathbf{y}] = \mathbf{f} + B\mathbf{z} \quad (4.21.1)$$

$$V[\Lambda|\mathbf{y}] = Q - K' H K \quad (4.21.2)$$

where

$$\begin{aligned} B &= \begin{pmatrix} I & 0 \\ 0 & H \end{pmatrix} \begin{pmatrix} I & 0 \\ -Q_{21} & V_2 \end{pmatrix} \\ H &= (Q_{22} + V_2)^{-1} \\ K &= (Q_{21} \quad Q_{22}) \end{aligned}$$

and  $\mathbf{z}$  is the vector of the  $(f_i^* - f_i)/q_i^*$ .

We should note, from (4.21.2), that the posterior covariance matrix  $V[\Lambda|\mathbf{y}]$  may depend on the new observation, since it depends on  $V_2$ . For example, if  $f(\cdot)$  links the gamma and log-normal distributions through coefficients of variation, this will be the case. Therefore, this model differs from the classical dynamic linear model, in the sense that we cannot obtain, a priori, the posterior variances.

#### 4.8 The general updating

Let  $k$  denote, as usual, the number of periods our processes last. We shall define a vector  $\Psi_t$ , by:

$$\Psi'_t = (\theta'_t, \theta'_{t+1}, \dots, \theta'_{t+k-1}) \quad (4.22)$$

where the vector space  $\theta_t$  evolves on time according to (4.6). Then, we can determine the distribution of  $(\Psi_t|D_{t-1})$ , if (4.7) is given. Let

$$(\Psi_t|D_{t-1}) \sim N[\alpha_t, \Sigma_t]$$

where  $\alpha_t$  and  $\Sigma_t$  are easily expressable in terms of  $a_t$ ,  $R_t$  and  $\delta$ .

Now, consider  $\Lambda_t$ , defined by:

$$\Lambda'_t = (\lambda'_t, \lambda'_{t+1}, \dots, \lambda'_{t+k-1})$$

where the  $\lambda$ 's are obtained as in (4.7). Then, we can write:

$$\Lambda_t = \Phi_t \Psi_t$$

where the matrix  $\Phi_t$  is given by:

$$\Phi_t = \begin{pmatrix} F'_t & 0 & \dots & 0 \\ 0 & F'_{t+1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & F'_{t+k-1} \end{pmatrix}$$

Now, we have prior distribution for  $\Lambda_t$ , in the form:

$$(\Lambda_t | D_{t-1}) \sim N[\mathbf{f}_t, Q_t]$$

where

$$\begin{aligned} \mathbf{f}_t &= \Phi_t \alpha_t \\ Q_t &= \Phi_t \Sigma_t \Phi'_t \end{aligned}$$

Apply the updating procedure just described, using the equivalent observation vector together with its associated variance matrix, to obtain:

$$(\Lambda_t | D_t) \sim N[\mathbf{f}_t^*, Q_t^*]$$

where  $\mathbf{f}_t^*$  and  $Q_t^*$  are obtained using (4.21.1) and (4.21.2), respectively. From this we can obtain  $(\Psi_t | D_t)$ :

$$(\Psi_t | D_t) \sim N[\mu_t, \Gamma_t]$$

where  $\mu_t$  and  $\Gamma_t$  are calculated using the same idea of (4.10.1) and (4.10.2):

$$\begin{aligned} \mu_t &= E[\Psi_t | D_t] = E[E\{\Psi_t | \Lambda_t, D_{t-1}\} | D_t] \\ &= E[\alpha_t + \Sigma_t \Phi'_t Q_t^{-1} (\Lambda_t - \mathbf{f}_t) | D_t] \\ &= \alpha_t + \Sigma_t \Phi'_t Q_t^{-1} (\mathbf{f}_t^* - \mathbf{f}_t) \end{aligned}$$

and

$$\begin{aligned} \Gamma_t &= V[\Psi_t | D_t] = V[E\{\Psi_t | \Lambda_t, D_{t-1}\} | D_t] + E[V\{\Psi_t | \Lambda_t, D_{t-1}\} | D_t] \\ &= V[\alpha_t + \Sigma_t \Phi'_t Q_t^{-1} (\Lambda_t - \mathbf{f}_t) | D_t] + E[\Sigma_t - \Sigma_t \Phi'_t Q_t^{-1} \Phi_t \Sigma_t | D_t] \\ &= \Sigma_t \Phi'_t Q_t^{-1} Q_t^* Q_t^{-1} \Phi_t \Sigma_t + \Sigma_t - \Sigma_t \Phi'_t Q_t^{-1} \Phi_t \Sigma_t \end{aligned}$$

Defining  $\mathbf{A}_t = \Sigma_t \Phi'_t Q_t^{-1}$ , we can write:

$$\begin{aligned} \mu_t &= \alpha_t + \mathbf{A}_t (\mathbf{f}_t^* - \mathbf{f}_t) \\ \Gamma_t &= \Sigma_t - \mathbf{A}_t (Q_t - Q_t^*) \mathbf{A}_t' \end{aligned}$$

The updating is closed by obtaining  $(\Psi_{t+1}|D_t)$ . We observe the definition of  $\Psi$  in (4.22). Let  $\mu_t$  and  $\Gamma_t$  be given as

$$\mu'_t = (\mu'_{t,1}, \mu'_{t,2}, \mu'_{t,3}, \dots, \mu'_{t,k})$$

$$\Gamma_t = \begin{pmatrix} \Gamma_{t,11} & \Gamma_{t,12} & \Gamma_{t,13} & \dots & \Gamma_{t,1k} \\ \Gamma_{t,21} & \Gamma_{t,22} & \Gamma_{t,23} & \dots & \Gamma_{t,2k} \\ \Gamma_{t,31} & \Gamma_{t,32} & \Gamma_{t,33} & \dots & \Gamma_{t,3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Gamma_{t,k1} & \Gamma_{t,k2} & \Gamma_{t,k3} & \dots & \Gamma_{t,kk} \end{pmatrix}$$

where the  $\mu_{t,i}$ 's are  $n$  dimensional vectors and the  $\Gamma_{t,ij}$  are  $n \times n$  matrices,  $n$  being the dimension of  $\theta$ . Then, we will have  $(\Psi_{t+1}|D_t) \sim N[\alpha_{t+1}, \Sigma_{t+1}]$ , where

$$\alpha_{t+1} = (\mu'_{t,2}, \mu'_{t,3}, \dots, \mu'_{t,k}, \mu'_{t,k})$$

$$\Sigma_{t+1} = \begin{pmatrix} \Gamma_{t,22} & \Gamma_{t,23} & \dots & \Gamma_{t,2k} & \Gamma_{t,2k} \\ \Gamma_{t,32} & \Gamma_{t,33} & \dots & \Gamma_{t,3k} & \Gamma_{t,3k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Gamma_{t,k2} & \Gamma_{t,k3} & \dots & \Gamma_{t,kk} & \Gamma_{t,kk} \\ \Gamma_{t,k2} & \Gamma_{t,k3} & \dots & \Gamma_{t,kk} & \Gamma_t^* \end{pmatrix}$$

where  $\Gamma_t^*$  will be obtained from  $\Gamma_{t,kk}$  by the discount factor approach. This puts the updating algorithm in closed form.

#### 4.9 Simulation results

We tested a simplified version of our model using simulated data, which were generated as follows. We choose a value  $\lambda_0$  and a set of  $k$  positive numbers  $r_1, r_2, \dots, r_k$  which sum to one. We also choose  $V > 0$  and generate  $p$  independent observations  $\omega_1, \omega_2, \dots, \omega_p$  of a  $N(0, V)$  distributed random variable and calculate

$$\varphi_j = \exp \left\{ \lambda_0 + \sum_{i=1}^j \omega_i \right\}, \quad j = 1, \dots, p$$

Then, we generate a random matrix  $X$ , where  $x_{ij}$  is an observation from a Poisson distribution with mean  $\varphi_i r_j$ . We now take this data and try to fit a simplified version of our model, in the sense that  $\theta$  is a univariate parameter and all the  $F$ 's are equal to one, which makes  $\theta$  coincident with  $\lambda$ . The fitted model is then slightly different from the 'true' model in the sense that the Poisson parameters are log-normal random variables, while, in our fitted model, they come from gamma distributions which are linked to log-normal ones through a known bijection.

For verification of model adequacy for the  $j$ -th week we choose the goodness of fit statistics given by

$$\frac{\sum_{i=1}^p (\log \pi_{ij}(x_{ij}) - \mu_{ij})}{(\sum_{i=1}^p \sigma_{ij}^2)^{1/2}} \quad (4.23)$$



where  $\pi_{ij}$  is the predictive distribution for  $x_{ij}$ ,  $x_{ij}$  is the actual observed value,  $\mu_{ij} = E[\log \pi_{ij}]$  and  $\sigma_{ij}^2 = V[\log \pi_{ij}]$ . We remember from (4.3) that the model fits a negative binomial distribution for  $x_{ij}$ , and, in our examples,  $\mu_{ij}$  and  $\sigma_{ij}$  are obtained numerically. The statistics obtained in (4.23) will be tested against a  $N(0, 1)$  distribution.

Before we discuss the simulation results, it may be important to understand the role played by the discount factor in the presented model. Let's consider the seven steps described at the beginning of Section 4.4. In the very particular case of our simulations,  $\lambda$  is coincident with  $\theta$ . This fact implies a reduction in that updating algorithm in the sense that the passage from step 2 to step 3, as well as that one from step 6 to step 7 simply become identities. Now, suppose we use a unit discount factor in the algorithm. Then, the distribution for  $\lambda_t$  obtained in step 6 will be exactly the distribution in step 3 of the next iteration. Consequently, for  $\delta = 1$ , the updating procedure is reduced to steps 4 and 5 only, which means we just use the classic conjugate method to update the gamma distribution of  $\varphi_t$ . This is equivalent to consider that the observations for the  $j$ -th period of the process, given a certain parameter  $\varphi$ , are generated by a Poisson distribution with mean  $r_j\varphi$  and the gamma distribution for  $\varphi$  is updated via conjugate analysis each time a new observation becomes available. We recall equation (4.4) to write the updated mean  $\mu^*$  of the gamma distribution in the form:

$$\mu^* = p\mu + qx \quad (4.24)$$

where  $\mu$  is the mean obtained in the last iteration,  $x$  is the value we just observed,  $p = \beta/(\beta + r)$  and  $p + q = 1$ . We can readily observe that, after a few iterations,  $p$  will be almost equal to one, which means that the fitted values will, in practice, fall on a horizontal straight line across the data. A discount factor equal to one represents a global approximation, in the sense that, in practice, we are using a constant model, except, perhaps, for the first few observations.

Now, suppose we use  $\delta < 1$ . From the expressions linking the mean and coefficient of variation of the log-normal with the mean and variance of the underlying normal (see Appendix) we can see that an increase in the underlying variance of the normal (which represents a discount factor smaller than one) implies an increase in the mean and coefficient of variation of the associated log-normal. Also note that these two quantities are exponential functions of the variance of the underlying normal, and, consequently, a reasonably small discount factor implies a considerable increase. Using (4.14.2) and (4.15) we can calculate the parameters of the associated gamma

distribution. Note that the  $\beta$  parameter of the gamma distribution is obtained as

$$\beta_G = \frac{1}{\mu_G k_G^2} \quad (4.25)$$

Using (4.15) this will be written as

$$\beta_G = \frac{0.2886 + k_G^{-2}}{\mu_L} \quad (4.26)$$

Therefore, if we increase  $\mu_L$  and  $k_L$ , then, by (4.26), we decrease  $\beta_G$  (observe that if we increase  $k_L$ , we increase  $k_G$ , inverting (4.14.2)). A reasonably small discount factor implies a considerable increase of  $\mu_L$ , since the growth is exponential. Therefore,  $\beta_G$  will become considerably small. Consequently, in the updating equation (4.24),  $p$  will become considerably small and the weight given to the observation will be much bigger than the one given to the prior mean. We conclude that the bijection we constructed using a quadratic relationship like (4.15) seems very convenient, since we continue to use the philosophy that a discount factor near one is translated as a simple model with an almost constant forecast function and the approximation becomes more local as we decrease the discount factor. Had we used a higher degree polynomial (a cubic relationship, for example), we would probably have problems, since the numerator in (4.26) would become an increasing function of  $k_G$ .

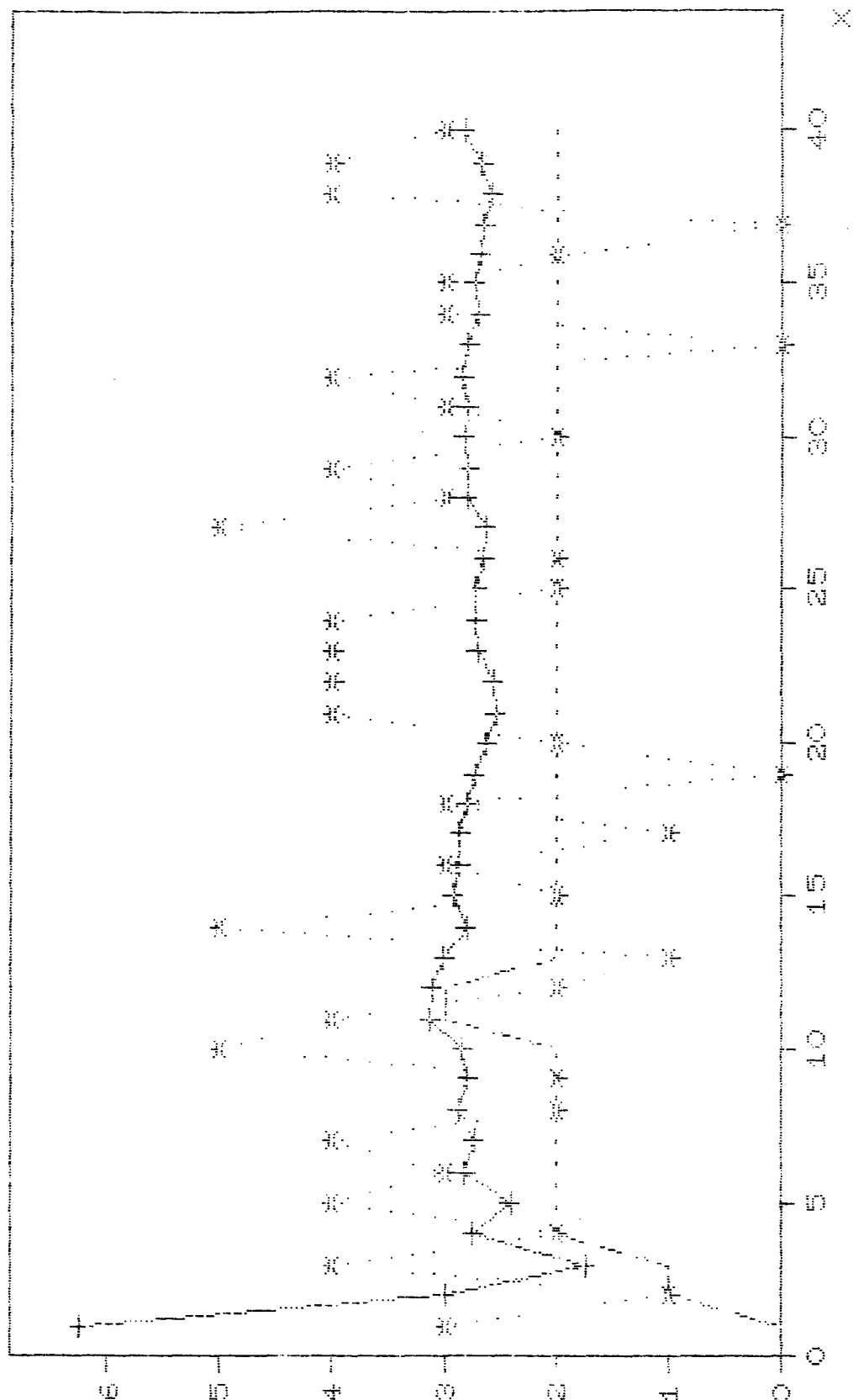
When we adopt the simple method of equating mean and coefficient of variation, it will be clear from (4.25) that a reasonable increase of mean and coefficient of variation represents a large decrease of  $\beta_G$ , as in the previous case.

We consider three simulations. For both of them the matrix  $X$  is  $40 \times 6$  and we choose  $\lambda_0 = 3$ . The reservation curve is the vector  $(0.2, 0.15, 0.1, 0.25, 0.15, 0.15)$ . In the first simulation, we have chosen  $V = 0.01$ , and the six series generated with this model do not show sudden variations of data. In the second one we have used  $V = 0.1$ , and the six series generated present much more variation. The third simulation uses  $V = 0.5$  (graphs can be seen in the next pages). For the first set of data, we fit a model with a discount factor  $\delta = 0.95$ . The result for the first period is presented in the first graph, (G 4.1), where we have plotted in the same graph the actual data and the mean and mode of the predictive distribution. The goodness of fit statistics given by (4.23) is shown in the bottom of the graph, and this must be tested against a  $N(0, 1)$  distribution. From the graph, we can see that the model seems to explain reasonably well the behaviour of data, and this can be achieved with the large discount factor of  $\delta = 0.95$ . The very high discount factor used in this model is more or less expected, since data were generated with a very small systematic variation ( $V=0.01$ ), and this means we are expecting to fit

an almost horizontal line through the points. The big discrepancy between the first observation and the first fitted point is due to the fact that we begin with a very uninformative prior for the normal distribution, and this implies a very high mean for the associated log-normal. Second plot (G 4.2) shows the first period of second set of series (with  $V = 0.1$ ) together with its mean and mode fits using the same discount factor  $\delta = 0.95$ . Observe that although the goodness of fit statistics looks fine for the fit, the errors still seem to present some pattern, almost all of them being negative. The third graph, (G 4.3), shows the performance of our model when we try to apply it to this same data, this time with a discount factor  $\delta = 0.8$ . We observe that this performance increases considerably, something that is also indicated by the goodness of fit statistics for the fit, the errors seeming to be more symmetrically distributed. This is clear from the fact that if data is subject to a higher level of variation, then, we must use a higher adaptive factor when trying to fit a model, that is, a smaller discount. Finally, the last graph (G 4.4) shows a simulation for which we have used  $V = 0.5$ . A reasonably good fit can be obtained with a discount factor of  $\delta = 0.75$ . The reasonably large goodness of fit statistics obtained here can be explained by the very big values of the seventh and eighteenth observations, which introduce a large forecasting error.

DATA (+) MEAN (x) MODE (.)

Y

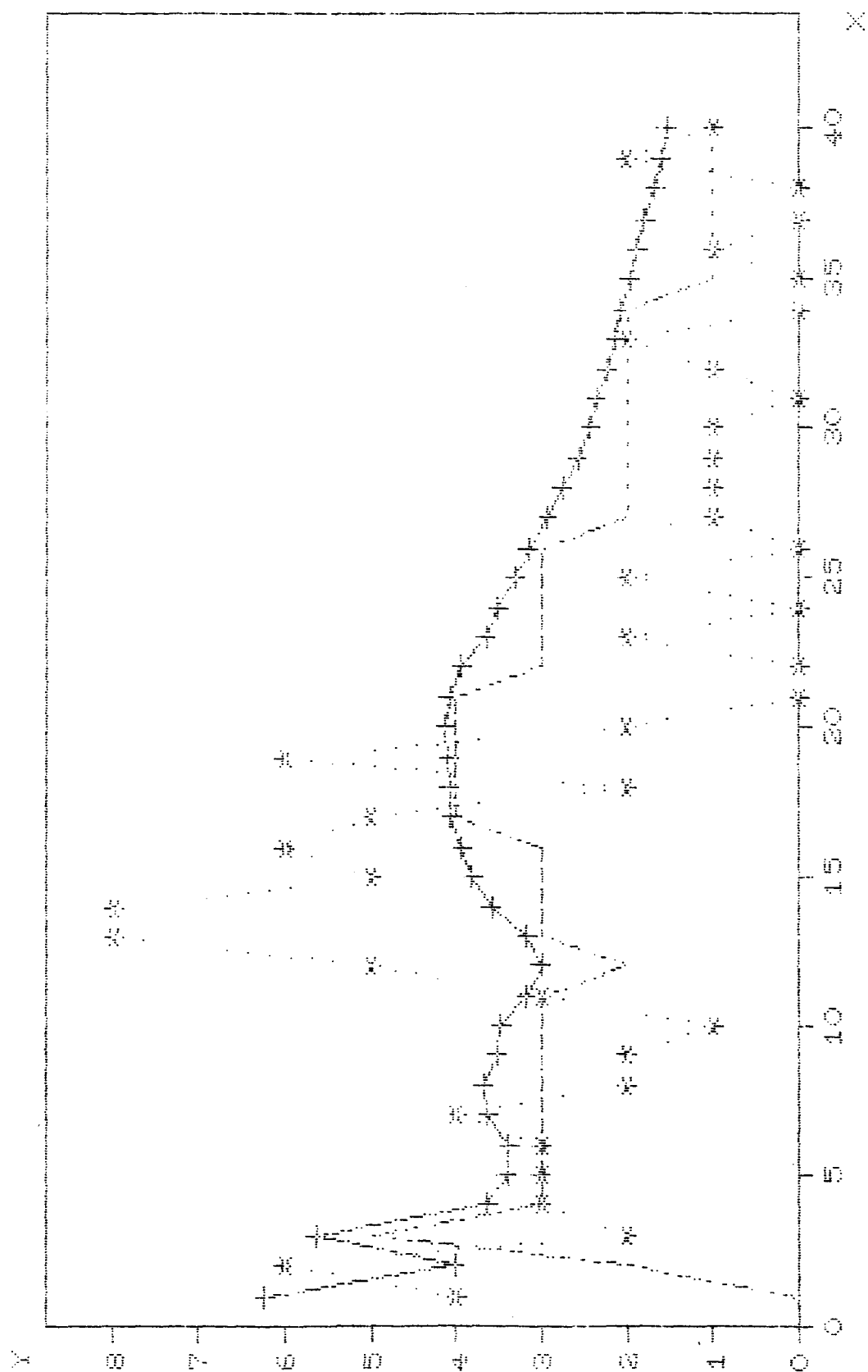


COEFFICIENT OF FIT: 0.676

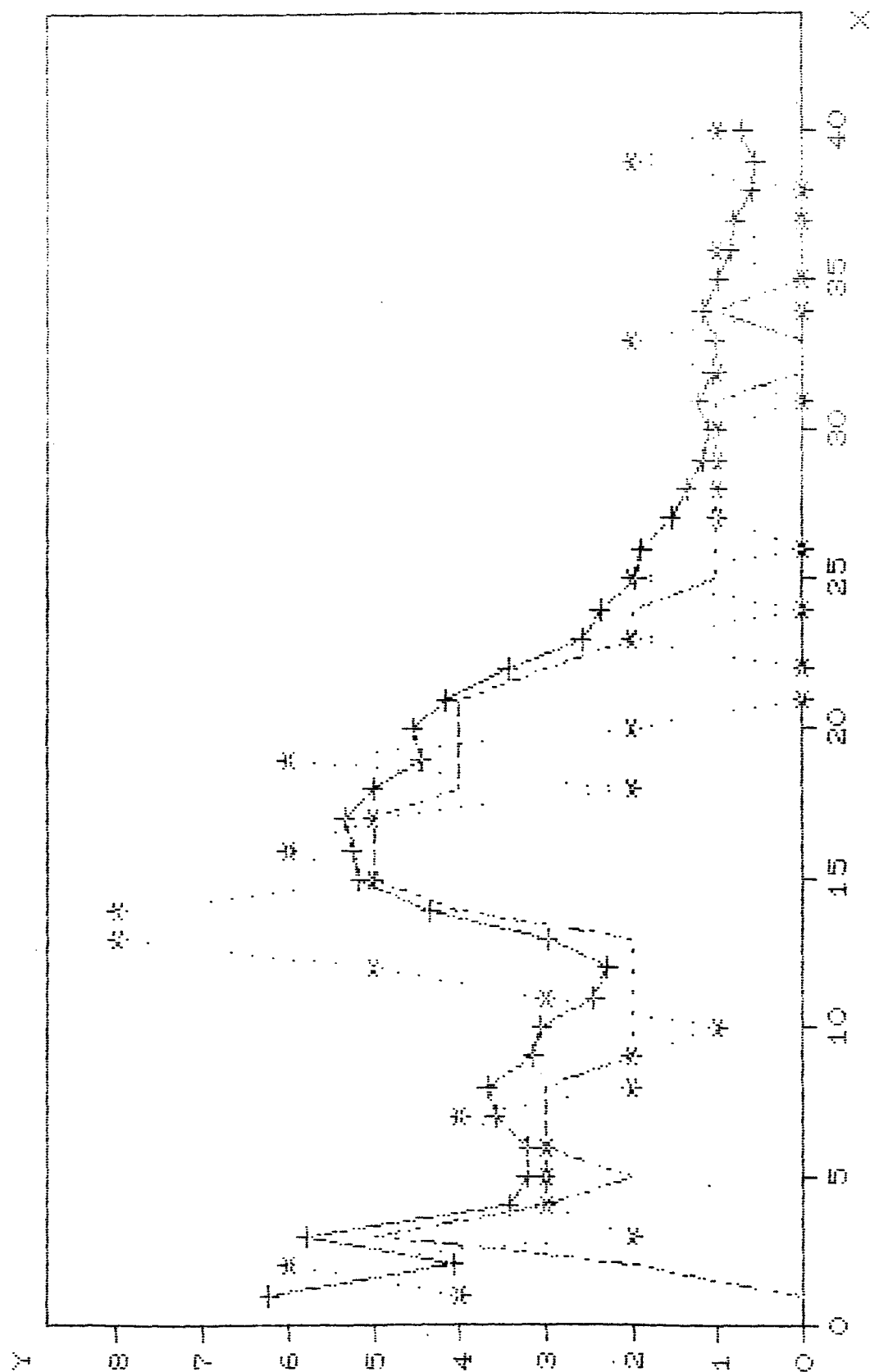
G 4.1

G 4.2

DATA (+) MEAN (•) MODE (.)

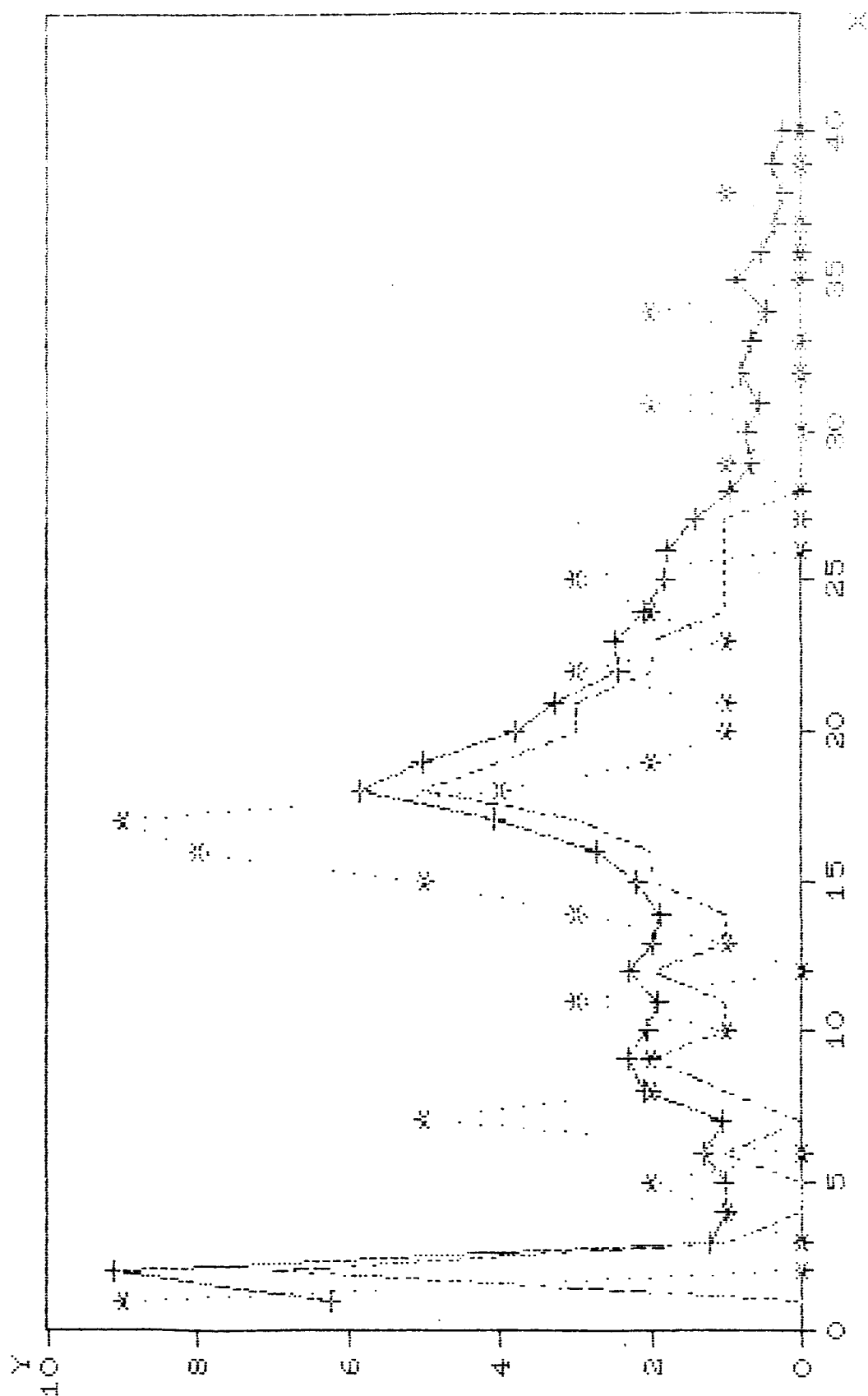


DATA (\*) MEAN (+) MODE (.)



G 4.3

DATA (+) MEAN (+) MODE (\*)



GOODNESS OF FIT: 72.188

G 4.4

#### 4.10 Conclusions

In all the graphs we observe that the mode of the forecast distribution is almost invariably equal to the integer part of its mean, that is, the mode is very near the mean. For a reasonably large discount factor, this result is expected, since the mode  $n^*$  of the negative binomial is obtained from its mean  $\mu$  through the expression

$$n^* = \max\left\{0, \left\lfloor \left\lfloor \mu - \frac{q}{p} \right\rfloor \right\rfloor \right\}$$

where  $\lfloor x \rfloor$  denotes the integer part of  $x$ , and  $p, q$  are as in (4.24). We remember that for a reasonably large discount factor, we have a reasonably large  $p$  and, consequently a small  $q/p$  ratio. From the expression above, we can clearly see, then, that the mode will, in most situations, be equal to the integer part of the mean. A small distance between these two parameters means that the spot decision represented by the mode (take the most probable value) is coincident to the least squares loss decision represented by the mean. From the fact that we are working with small integer numbers (not greater than ten), the expected value being almost equal to the mode implies that our forecast distributions are quite concentrated in the smallest integers, the probability distribution function presenting a quick fall (small after ten), which implies a small variance of the forecast distribution.

Another important conclusion from the results we obtained is that the log-normal to gamma density approximation seems to do a good job when used to define the bijection in our model (note that data are generated directly from a log-normal distribution and the model supposes a gamma distribution for the Poisson parameters).

We must observe that, when dealing with larger variations of data, we are expected to work with smaller discount factors. For a reasonably high level of variation, we should work, in principle, with a reasonably small discount factor, and, for large variations we are forced to adopt a very small discount factor. This will not represent a good model, since we will be virtually repeating the last observation. An alternative approach is try to smooth out the very high frequencies presented in the series, for example, by data aggregation. This approach will lead us to the study of the aggregation of observations, which constitutes the main topic of the following chapter.



#### 4.11 Appendix

We consider the problem of approximating a gamma distribution by a log-normal and vice-versa. In this approach, we consider two methods. The first one is the simple method which consists of equating mean and variance for both distributions. The second one is an attempt to find a density having the smallest  $L_2$  distance to the given density.

We divide the appendix to this chapter in four parts. In the first part, we present a summary of the features for both distributions. In the second part, we consider the method of equating mean and variance for both distributions, giving some examples. In the third part, the method consisting of minimization of the  $L_2$  distance is explained. We derive a guide relationship, to obtain a log-normal from a gamma and vice-versa, based on the results obtained with this method, and verify that this guide relationship seems to work very well for small values of the coefficient of variation of the distribution. The fourth part is a comparison of both methods, with conclusions.

##### PART 1. *Basic properties of both distributions*

###### 1) Log-normal

Density is defined over  $\mathbf{R}^+$  as:

$$f(x|m, s^2) = \frac{1}{sx\sqrt{2\pi}} \exp \left\{ \frac{-(m - \log x)^2}{2s^2} \right\} \quad (4.27)$$

If  $X$  is a random variable with density given by (4.27):

$$\begin{aligned} \mu &= E[X] = \exp\{m + \frac{1}{2}s^2\} \\ \sigma^2 &= Var[X] = \mu^2(\exp(s^2) - 1) \end{aligned}$$

Let  $k$  be the coefficient of variation and define the quantity  $a = \sqrt{1 + k^2}$ . We get:

$$\begin{aligned} s^2 &= 2 \log a \\ m &= \log \left( \frac{\mu}{a} \right) \end{aligned}$$

Also, the median  $x_m$  of the distribution is  $x_m = e^m$  or

$$x_m = \frac{\mu}{a}$$

The mode  $x^*$  of the distribution is

$$x^* = \exp\{m - s^2\} = \frac{\mu}{a^3}$$

The density function is unimodal and always has two inflexion points, given by:

$$x_c = \frac{x^*}{a} \exp \left\{ \frac{\pm s \sqrt{s^2 + 4}}{2} \right\} \quad (4.28)$$

Hence, using the approximation  $e^x \approx 1 + x$ , when  $x$  is small, we can see that for a small value of  $k$ , the inflexion points are approximately symmetric around the median (mode) of the distribution.

## 2) Gamma

Density is defined over  $\mathbb{R}^+$  as:

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad (4.29)$$

Let  $X$  be a random variable with this density. Then, if  $\mu = E[X]$  and  $k$  is the coefficient of variation, we have:

$$\alpha = \frac{1}{k^2} \quad (4.30)$$

$$\beta = \frac{1}{\mu k^2}$$

Mode of distribution does not exist if  $\alpha < 1$ . If  $\alpha \geq 1$ , the mode  $x^*$  is given by:

$$x^* = \mu(1 - k^2)$$

If  $\alpha \leq 1$ , the density curve has no inflexion points, the curve tending to infinity, as  $x$  approaches zero, for  $\alpha < 1$ . If  $1 < \alpha \leq 2$ , the curve has a single inflexion point  $x_c$ , given by:

$$x_c = x^* + k\sqrt{\mu x^*}$$

If  $\alpha > 2$ , the curve has two inflexion points, which are symmetric around the mode, and are given by:

$$x^* \pm k\sqrt{\mu x^*}$$

## PART 2. Equating mean and variance

A first attempt to approximate a gamma distribution by a log-normal and vice-versa consists of equating mean and variance for both distributions. We should expect this procedure to be fairly good, if the coefficient of variation  $k$  is relatively small. In fact, if  $k$  is small, we must have a large value of  $\alpha$  for the corresponding gamma distribution, by (4.30). This gamma distribution with a large value of  $\alpha$  can be seen as the distribution of the sum of many independent gamma distributions. Then, because of the central limit theorem we can say this distribution may be well approximated by a normal distribution.

Also, if  $X$  is a log-normal random variable with density given by (4.27) and  $\mu = E[X]$ , we can rewrite (4.27) as

$$f(x|\mu, s^2) = \frac{1}{s\mu\sqrt{2\pi}} \exp \left\{ s^2 - \frac{1}{2} \left( \frac{1}{s} \log \left( \frac{x}{\mu} \right) + \frac{3}{2}s \right)^2 \right\} \quad (4.31)$$

Now, when  $x$  is very close to  $\mu$ , we can use the approximation  $\log r \approx r - 1$ , if  $r \approx 1$ . From the same approximation, we can see that for a very small coefficient of variation  $k$  of the log-normal, we will have  $k \approx s$ , which means that  $s\mu$  will be approximately the standard deviation  $\sigma$  of the log-normal distribution. This gives an approximation of (4.31) by:

$$f(x|\mu, \sigma) \approx \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\}$$

which is the density of a normal distribution. We can also see, by (4.28), that if the coefficient of variation is very small, the inflection points will be very close to the median (which will be also approximately the mean) of the distribution.

Hence, we can see that if the coefficient of variation is small, both curves can be very well approximated by a normal curve, and must be very close to each other.

On the other hand, if the coefficient of variation is large, we must expect the curves to differ. For example, we can see by (4.30) that if  $0.5 \leq k^2 < 1$  the gamma distribution will have only one inflection point, while the log-normal density always has two inflection points. Also, for  $k \geq 1$ , the two densities will be completely different. For example, we can easily see by (4.27) that the density  $f$  of a log-normal distribution necessarily satisfies  $f(0^+) = 0$ . The density  $g$  of a gamma distribution is, by (4.29), such that  $g(0^+) = \beta > 0$ , if  $k = 1$ , and  $g(0^+) = +\infty$ , if  $k > 1$ .

We have used the approach of equating mean and variance, fixing the mean equal to one, and varying the coefficient of variation  $k$  from 0.1 to 1.1, in steps of 0.1. For  $k = 0.1$  and 0.2 the two densities are very similar to each other and the normal density can be a good approximation to them. If we consider larger values of  $k$ , the curves begin to differ as  $k$  increases; the log-normal density always shows a higher peak. Also, we can see that for reasonably high values of  $k$  the gamma distribution is more heavy in a small neighbourhood of zero. For  $k = 0.5$  there is a considerable difference between the two curves, which justifies a search for another approximation method.

### PART 3. $L_2$ minimization

Because equating the first two moments does not produce satisfactory results when we consider a relatively large value of  $k$ , we can think about trying to use another

procedure for approximation. One idea is try to minimize the  $L_2$  distance between the densities. In other words, if  $f$  is the density we want to approximate by a curve belonging to a specific family, we try to find the curve  $g^*$  of this family such that

$$d_{L_2}(f, g) = \int_0^\infty |f(x) - g(x)|^2 dx \quad (4.32)$$

is minimized for all  $g$  in the family.

Now, suppose  $f$  is the gamma density defined by (4.29) and  $g$  is the log-normal density given by (4.27). Then, we can easily verify that  $d_{L_2}(f, g)$ , given by (4.32), is finite if and only if  $\alpha > 1/2$  and, in this case, (4.32) becomes:

$$\frac{\beta\Gamma(2\alpha-1)}{2^{2\alpha-1}\Gamma^2(\alpha)} + \frac{1}{2s\sqrt{\pi}} \exp\left\{\frac{s^2}{4} - m\right\} - \frac{\beta\sqrt{2}}{s\Gamma(\alpha)\sqrt{\pi}} \int_0^\infty h(u) du \quad (4.33)$$

where  $h(u)$  is given by:

$$h(u) = e^{-u} u^{\alpha-2} \exp\left\{\frac{-(m + \log \beta - \log u)^2}{2s^2}\right\} \quad (4.34)$$

Then, for example, to approximate a gamma density with  $\alpha > 1/2$  by a log-normal density, a possible approach is try to minimize (4.33) with respect to  $m$  and  $s$ , given the fixed values of  $\alpha$  and  $\beta$ . Because (4.32) becomes a complicated function of  $m$  and  $s$ , we must, in practice, use a numerical approach, which means perform minimization using a nonlinear programming technique. We have tried some examples using the Fletcher-Reeves version of the conjugate gradient method. The conjugate gradient was chosen because of its relatively good efficiency and the F.R. version was used because it can guarantee global convergence of the algorithm. As this is an unconstrained method and we necessarily have  $s > 0$ , we had to consider a change of parameters introducing a new parameter  $p$  such that  $s = e^p$ . Hence, our aim is to find  $m$  and  $p$  such that (4.33) is minimized,  $m$  and  $p$  unconstrained.

To implement the numerical optimization using this method, we also need the derivatives of the function  $F(m, p)$  to be minimized; in this case,  $F(m, p) = d_{L_2}(f, g)$ , given by (4.33), for which the partial derivatives with respect to  $m$  and  $p$  are:

$$\frac{\partial F}{\partial m} = \frac{-1}{2s} \exp\left\{\frac{s^2}{4} - m\right\} + \frac{\beta\sqrt{2}}{s^2\Gamma(\alpha)} \int_0^\infty R(u)h(u) du \quad (4.35)$$

$$\frac{\partial F}{\partial p} = \left(\frac{s}{4} - \frac{1}{2s}\right) \exp\left\{\frac{s^2}{4} - m\right\} + \frac{\beta\sqrt{2}}{s\Gamma(\alpha)} \int_0^\infty (1 - R^2(u))h(u) du \quad (4.36)$$

where  $h(u)$  is given by (4.34) and

$$R(u) = \frac{m + \log \beta - \log u}{s} \quad (22)$$

The integrals appearing in (4.33), (4.35) and (4.36) may be evaluated numerically by Gaussian quadrature. Here, we may use the Laguerre polynomials to take advantage of the fact that we are integrating from zero to infinity and the function  $h(u)$ , appearing in the integrands, involves the factor  $e^{-u}$ .

We have used this optimization approach, trying to approximate the density of a gamma distribution with unit mean by a log-normal density. We considered values of the coefficient of variation  $k$  from 0.3 to 1.0, in steps of 0.1. For  $k$  running from 0.3 to 0.5 the log-normal density obtained by the method is very close to the original gamma density, the log-normal always having a higher peak. Also, both tails of the log-normal density are slightly shifted to the right, in comparison with those for the gamma. From  $k = 0.6$  there is a reasonable difference between the two densities. In fact, for  $k = 1$ , we know the configurations to be completely different from one curve to the other. However, because a large coefficient of variation means a large uncertainty about the data, the difference between the two densities does not constitute a big problem, for large  $k$ .

When trying to apply an optimization algorithm to minimize our distance, some problems may happen. One of the most common problems comes from the fact that the function may have more than one point of local minimum. Hence, the answer we find by using the method may depend on the initial point. Also, for some special cases of the density we want to approximate, we can have problems with the Gaussian integration. For example, when trying to approximate a log-normal by a gamma density,  $\alpha$  may become very large in a certain step of the algorithm, which can cause a numerical overflow. Moreover, the algorithm may take a lot of time for convergence, depending on the initial value and the behaviour of the objective function. Thus, it makes sense to consider an approximate analytic relation between the parameters of the density to be approximated and those for the best approximating density. We have obtained this approximate relation by running the algorithm for twelve different values of  $k$ , between 0.275 and 0.525 and a fixed unit mean for the gamma distribution, so that  $\mu_G = 1$ . By observing the coefficient of variation  $k_L$  and the mean  $\mu_L$  of the best log-normal obtained in each case, we derived an approximate relation between  $k_L$  and  $k$ , and also between  $\mu_L$  and  $k$ , through a linear regression. The two performed regressions yields the following equations:

$$k_L = 0.9135k + 0.4477k^2 \quad (4.37)$$

$$\mu_L = 1 + 0.2886k^2 \quad (4.38)$$

where  $k$  is the coefficient of variation of the gamma distribution we want to approximate. Relation (4.37) was obtained by fitting a quadratic polynomial passing

through the origin, while (4.38) was obtained by fitting a quadratic polynomial with a unit intercept and a null linear term, and estimating the quadratic coefficient. Regression (4.37) gives  $R^2 = 0.98$ , the error term having an estimated variance of  $2 \times 10^{-4}$  and an estimated first-order autocorrelation of  $-0.17$ . Regression (4.38) gives  $R^2 = 0.89$ , the error term having an estimated variance of  $6 \times 10^{-5}$  and an estimated first-order autocorrelation of  $0.15$ . These results show a very good fitting and we can use (4.37) and (4.38) to approximate a gamma with a unit mean.

Now, if  $X$  has a gamma distribution and  $Y$  has a log-normal distribution, then, for any  $c > 0$ , we know that  $cX$  has a gamma distribution and  $cY$  has a log-normal distribution. Using this, (4.38) can be easily generalized in order to approximate a gamma density with arbitrary mean by a log-normal density. Given a gamma distribution with mean  $\mu_G$  and coefficient of variation  $k_G$ , we obtain the mean  $\mu_L$  and coefficient of variation  $k_L$  for the approximating log-normal via the following guide relations:

$$\mu_L = \mu_G(1 + 0.2886k_G^2) \quad (4.39)$$

$$k_L = 0.9135k_G + 0.4477k_G^2 \quad (4.40)$$

The inverse relation may now be used to approximate a log-normal density by a gamma. Given  $\mu_L$  and  $k_L$  we may use (4.40) to obtain  $k_G$ , and from (4.39) we have  $\mu_G$ .

We have tried the guide relations above to approximate the two densities. We can observe the results obtained are as good as those obtained by using directly the optimization approach, not only for the interval in which we considered the various values of  $k_G$  in the regression, but also when we extrapolate the relation for values of  $k_G$  outside this interval. This leads us to the conclusion that the guide relations (4.39) and (4.40) can be considered a good solution to our approximation problem.

#### PART 4. *Comparison of the two methods and conclusions*

We consider the approximation of a log-normal by a gamma density. For small values of the coefficient of variation  $k_L$ , ( $k_L < 0.25$ ), the two methods show equally good performance. This can be justified by the fact that when  $k_L$  is small both curves can be very well approximated by a normal density. For larger values of  $k_L$  the guide relationship has a better performance than the simple method of equating the first two moments of the distributions. For  $k_L$  between 0.3 and 0.5 the guide relation produces a very good approximation (also true for  $k_G$  in the same interval). For the same values of  $k_L$ , the two curves differ significantly when we use the most

simple method. It can be seen that the performance of the second method is fairly better for values of  $k_L$  greater or equal to 0.6.

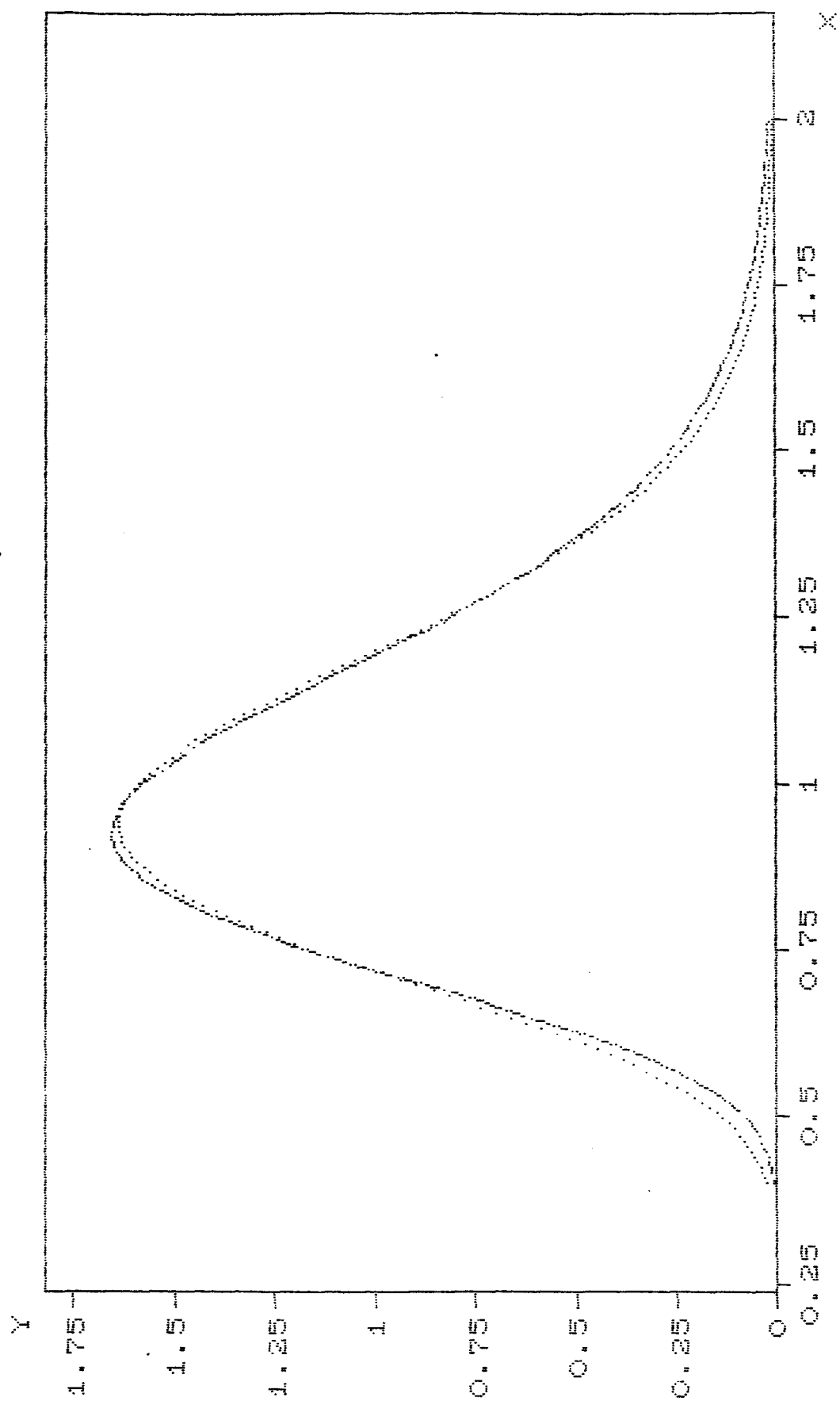
The graphs (G 4.5), (G 4.6) and (G 4.7) in the next three pages give the approximation for  $k_G$  equal to 0.25, 0.5 and 0.8, respectively. When approximating the two densities using (4.39) and (4.40), we can see that the log-normal always shows higher peak, mean and variance than the gamma distribution. The gamma density is slightly more shifted to the left, which means it accepts more easily the outcome of values near zero. The coefficient of variation of the log-normal is also greater than the one for the gamma. Hence, if we consider our method as an association rule between the two curves, we conclude that the log-normal associated with the gamma brings a bigger level of dispersion. We should note as well that the use of (4.32) as a measure of distance is indeed a good choice, specially for small values of the coefficient of variation of the curve we want to approximate. Indeed, when  $k$  is small, we can verify the two curves are very close to each other in the much stronger sense that

$$\sup_{x>0} |f(x) - g(x)| \tag{4.41}$$

is relatively small. So the  $L_2$  minimization, for small  $k$ , can be considered almost as good as minimization of (4.41), which is a much stronger approach.

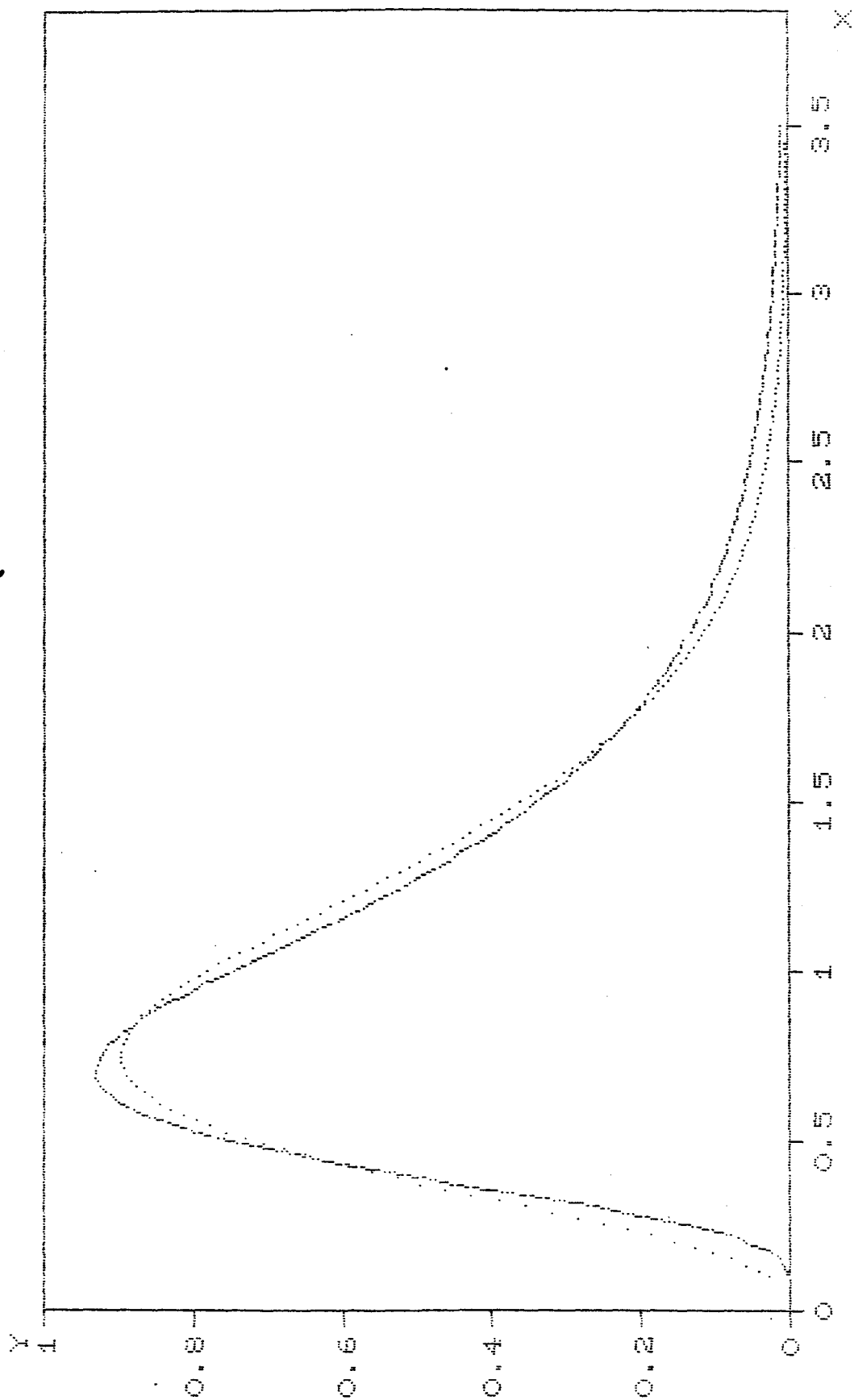
G 4.5

GAMMA (.) LOG-NORMAL (-) .



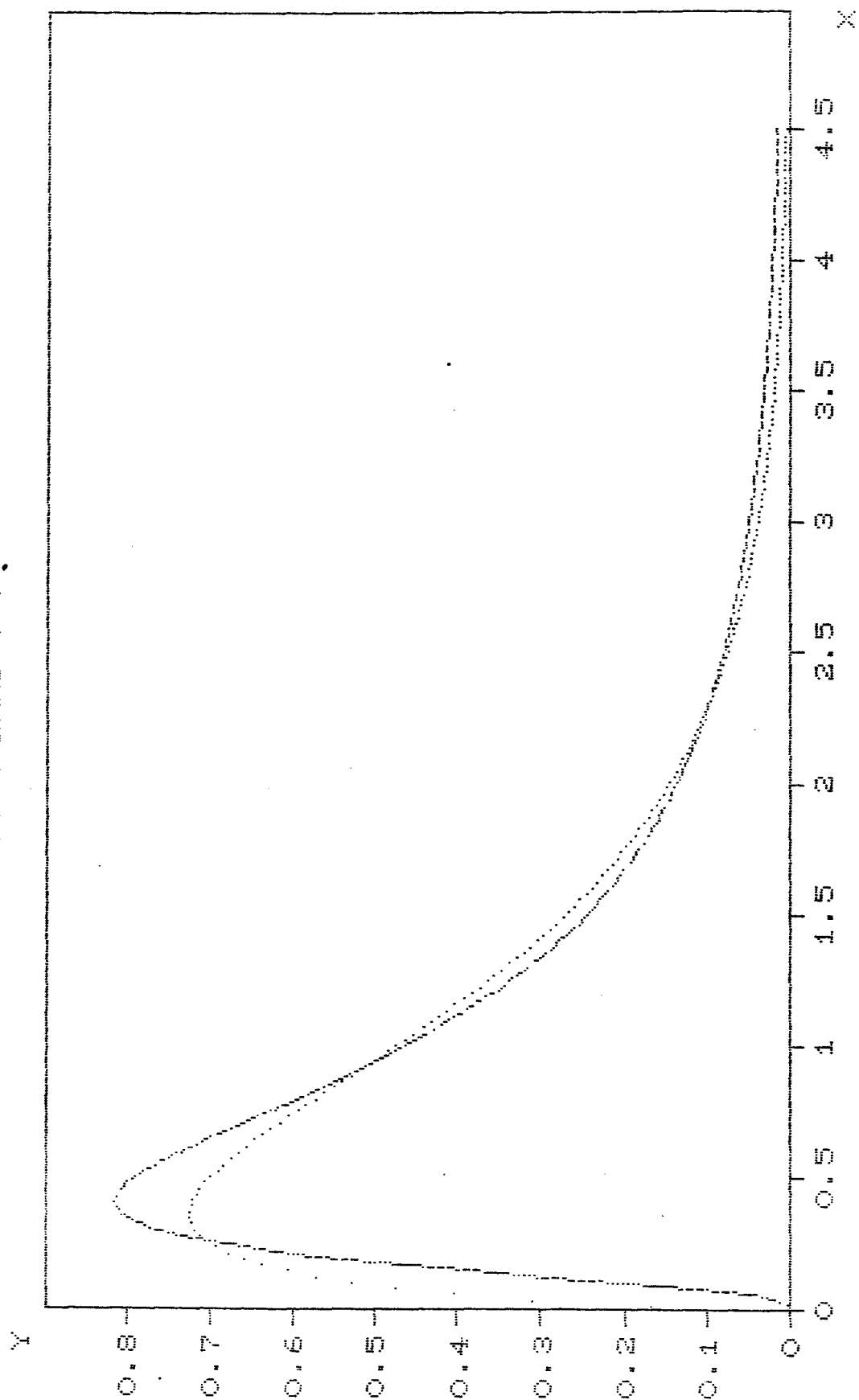


GAMMA (.) LOG-NORMAL (-).



G 4.6

GAMMA (.) LOG-NORMAL (-).



G 4.7

## CHAPTER 5

### DATA AGGREGATION

#### 5.1 Introduction

One important aspect concerning the analysis and forecasting of time series that is frequently neglected is the relationship between a model and the sampling interval, in particular, when the observation is cumulative over the sampling period. In other words, suppose we can observe a certain set of variables  $\{Y_t\}$  each unit of time. Now, we can collect the observations of our process each  $n$  units of time, defining

$$Z_k = \sum_{i=1}^n Y_{(k-1)n+i} \quad (5.1)$$

The relevant factors explaining the variation of  $Z_k$  can, and in general will, be different, depending on how we choose the sampling interval, i.e., on how we choose  $n$ . Consider the following example, given by Green and Harrison (1972). Suppose we have 1000 variables  $\{Y_i\}$  which can be decomposed in

$$Y_i = f_i + X + \epsilon_i,$$

for  $i = 1, \dots, 1000$ , where  $f_i$  is the known expected value of  $Y_i$ ,  $X$  is the common factor independent of all the  $\epsilon_i$ 's and the  $\epsilon_i$ 's are i.i.d. random variables. If  $Var[X] = 1$  and  $Var[\epsilon_i] = 99$ , for all  $i$ , then, we can obviously state that the effect of  $X$  in explaining  $Y_i$  is very small in comparison with  $\epsilon_i$ , which plays a much more important role in the model. But, if we now aggregate the variables, we have

$$Y = \sum_{i=1}^{1000} Y_i = \sum_{i=1}^{1000} (f_i + X + \epsilon_i) = f + 1000X + \epsilon,$$

where

$$f = \sum_{i=1}^{1000} f_i \quad \text{and} \quad \epsilon = \sum_{i=1}^{1000} \epsilon_i$$

Now, we will have  $Var[1000X] = 10^6$  and  $Var[\epsilon] = 99,000$ . Hence, the common factor  $X$ , which has very little importance for the simple model, will become a fundamental factor for explanation of the sum. It is very clear, therefore, that a simple model which can be suitable for a pre-specified sampling level can be inappropriate for another level, needing revision and probable sophistication.

Similarly, we can think of these aspects when the set of  $Y_i$ 's represents the observations of a process which can be measured for each unit of time; and we think of observing the process only after each  $n$  instants. For example, we could allow for

the  $\{\epsilon_i\}$  in the situation above to present some correlation structure, and the effect of aggregation becomes more important if these variables are negatively correlated, resulting in ‘cancelation’ effects. Also, aggregation of many independent and completely random effects can result in a total effect which may be significant for the new model. It is easy to see that if  $X \sim N(0, 1)$  and  $Y$  has a chaotic distribution, then we will not identify the distribution of  $Z = X - Y$ , but  $Y$  and  $Z$  sum up to produce a very simple  $N(0, 1)$  distribution.

On speaking of DLM models we can give very simple examples when aggregation can produce a great level of uncertainty and model revision should be considered. For example, suppose we are given a set of data which can be very well modelled by the simple univariate random walk + noise process, with equations:

$$\begin{cases} y_t = \theta_t + v_t & (5.2) \\ \theta_t = \theta_{t-1} + \omega_t, & (5.3) \end{cases}$$

where  $Var[v_t] = V$  and  $Var[\omega_t] = W$ , with known  $V$  and  $W$ .

Although the time evolution above can be a very good choice for modelling  $\{y_t\}$ , the aggregation of this data will possibly produce a lot of uncertainty, and the aggregated series will perhaps be better explained by a more complex structure. We should allow for example, for a correlation between  $\theta_t$  and  $v_t$ , in order to explain the bigger variability of data. It is always good to bear in mind that models will try to provide a reasonably good explanation for *local* behaviour of data, and the utility of a time series model is necessarily linked to the extent of time we consider for trying to explain data evolution.

In this chapter we will study aggregation of time series and how the aggregation of observations can possibly influence the forecasting performance. We begin by studying the very simple constant first order polynomial DLM and find some conditions for which the aggregated series can be represented by this same model. We also show how this model can be sophisticated when these conditions cannot possibly be satisfied, and verify that the sophisticated model can always give a representation for the aggregated data. We then study the aggregation of data, when the process follows the  $\{1, \lambda, V, W\}$  model. We show that the aggregated data can still be represented by a simple DLM model, provided some conditions are satisfied, and give a sophisticated DLM model that can always be used to represent the aggregated series. We discuss the importance of this sophisticated model and how the use of the more simple model affects the forecasting. The general problem (that of a DLM representation for the aggregated series in a general TSDLM model) is briefly discussed as is the linear growth model. Then, we turn to the model of the last

chapter, and discuss how aggregation influence the performance of the forecast for the booking of flights.

## 5.2 The constant first order polynomial DLM $\{1, 1, V, W\}$ case

### 5.2.1 The $\{1, 1, V^*, W^*\}$ representation

For the constant  $\{F, G, V, W\}$  DLM, we would like to know if the aggregated data can be represented by a similar DLM, and, if not, if there is a DLM representation for the aggregated series. We begin by considering the very simple constant first order polynomial DLM above and try to extend our results for more sophisticated processes.

**THEOREM 1 (DLM REPRESENTATION FOR AGGREGATED DATA  $\{1, 1, V, W\}$  CASE).** Suppose we are given a process  $\{y_t\}$ , evolving in time according to the univariate  $\{1, 1, V, W\}$  model with initial information  $(\theta_0|D_0) \sim N[m_0, C_0]$ , where  $\theta_t$  is the state parameter. Consider the new series  $\{Z_k\}$ , obtained from the first by aggregating each  $n$  observations, as in (5.1). If the conditions

$$\frac{V}{W} \geq \frac{n^2 - 1}{6} \quad \text{and} \quad \frac{C_0}{W} \geq \frac{n - 1}{2} \quad (5.4)$$

are satisfied, then the new aggregated series can be represented by the  $\{1, 1, V^*, W^*\}$  model

$$\begin{cases} Z_k = \phi_k + \epsilon_k & (5.5) \\ \phi_k = \phi_{k-1} + \delta_k, & (5.6) \end{cases}$$

with initial information  $(\phi_0|D_0) \sim N[\mu_0, \Gamma_0]$ , where

$$V^* = n\{V - \frac{n^2 - 1}{6}W\} \quad (5.7)$$

$$W^* = n^3W \quad (5.8)$$

$$\mu_0 = nm_0 \quad (5.9)$$

$$\Gamma_0 = n^2\{C_0 - \frac{n - 1}{2}W\} \quad (5.10)$$

**PROOF:** It suffices to show that expressions for the quantities  $E[Z_j|D_0]$ ,  $Var[Z_j|D_0]$  and  $Cov[Z_j, Z_k|D_0]$  calculated from the assumed model for  $y_t$  are the same as those calculated from the above model for  $Z_k$ , for all  $j, k$ . For simplicity, denote any  $F[X|D_0]$  (mean, variance or covariance) simply by  $F[X]$ . Clearly, (5.9) is satisfied, since:

$$E[Z_k] = nm_0 = \mu_0, \quad \forall k \quad (5.11)$$

hence, the expected values are the same in both cases.

Now, from (5.1), (5.2) and (5.3),  $Z_1$  can be rewritten as:

$$Z_1 = n\theta_0 + \sum_{i=1}^n (n+1-i)\omega_i + \sum_{i=1}^n v_i$$

and, therefore:

$$Var[Z_1] = n^2 Var[\theta_0] + nV + \frac{n(n+1)(2n+1)}{6}W \quad (5.12)$$

In general, we have:

$$Var[Z_k] = n^2 Var[\theta_{(k-1)n}] + nV + \frac{n(n+1)(2n+1)}{6}W$$

Using (5.3):

$$Var[\theta_{(k-1)n}] = C_0 + (k-1)nW \quad (5.13)$$

Hence:

$$\begin{aligned} Var[Z_k] &= n^2 C_0 + n^3(k-1)W + nV + \frac{n(n+1)(2n+1)}{6}W \\ &= n^2 C_0 + n^3 k W + nV - \frac{n(4n^2 - 3n - 1)}{6}W \end{aligned} \quad (5.14)$$

Now, using representation in (5.5) and (5.6):

$$Var[Z_k] = \Gamma_0 + kW^* + V^* \quad (5.15)$$

Substituting (5.7), (5.8) and (5.10) in (5.15) we arrive at (5.14).

It remains to calculate the covariance between  $Z_k$  and  $Z_{k+j}$ , for all  $k, j$ , using both forms. We first observe that because of (5.2) and (5.3),  $Z_k$ , defined by (5.1), can be written as:

$$Z_k = n\theta_{(k-1)n} + \sum_{i=1}^n (n+1-i)\omega_{(k-1)n+i} + \sum_{i=1}^n v_{(k-1)n+i} \quad (5.16)$$

Similarly, we have:

$$Z_{k+j} = n\theta_{(k+j-1)n} + \sum_{i=1}^n (n+1-i)\omega_{(k+j-1)n+i} + \sum_{i=1}^n v_{(k+j-1)n+i} \quad (5.17)$$

From (5.16) and (5.17) we see that  $Cov[Z_k, Z_{k+j}] = nCov[Z_k, \theta_{(k+j-1)n}]$ . From this and (5.3), we get:

$$\begin{aligned} Cov[Z_k, Z_{k+j}] &= n^2 Var[\theta_{(k-1)n}] + \frac{n^2(n+1)}{2}W \\ &= n^2 C_0 + n^3(k-1)W + \frac{n^2(n+1)}{2}W \end{aligned} \quad (5.18)$$

Also, from representation forms (5.5) and (5.6), we obtain:

$$Cov[Z_k, Z_{k+j}] = Var[\phi_k] = \Gamma_0 + kW^* \quad (5.19)$$

Using (5.8) and (5.10) in (5.19) we arrive at (5.18), and the result is proved.

From this theorem we can obtain an important consequence. Suppose, for example, that our original model is the  $\{1, 1, 1, 2\}$  constant model and we aggregate each  $n = 2$  observations. It can be readily seen that the aggregated series,  $Z_k$  can be represented by a simple random walk

$$Z_k = Z_{k-1} + u_k$$

where  $Var[u_k] = 16$ . This shows that, sometimes, a simple model can be obtained by aggregating observations from a more sophisticated model. Also, it is important to recall that, in the original formulation of the DLM, the observational and evolution errors are completely independent. Therefore, we could not expect, in principle, that these errors can be combined to produce a simple model where we only have systematic variation of data.

From (5.4) we can readily see that if we increase the level of aggregation, that same form of representation for the aggregated data can be maintained only up to a certain limit. After a certain point (i.e., a sufficiently large value of  $n$ ), we will violate the restrictions. That means we must look for a more sophisticated form of representing the evolution of the aggregated data. That can be interpreted by saying that after a certain reasonably large level of aggregation, we are led with too much data gathered together, which implies a loss of information given by the aggregated data. This loss of information will force us to model sophistication, leading us to introduce other parameters.

### 5.2.2 The sophisticated model

An idea for a more sophisticated model consists in the inclusion of a null eigenvalue in the system matrix. Equivalently, we allow for a correlation between the errors  $\epsilon_k$  and  $\delta_k$  in the observational and evolution equations, instead of supposing them to be independent. We consider, therefore, the following model for  $Z_k$ :

$$Z_k = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} \quad (5.20)$$

$$\begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_{k-1} \\ \epsilon_{k-1} \end{pmatrix} + \begin{pmatrix} \delta_k \\ \epsilon_k \end{pmatrix} \quad (5.21)$$

where

$$Var[\epsilon_k] = V^* \quad (5.22)$$

$$Var[\delta_k] = W^* \quad (5.23)$$

$$Cov[\epsilon_k, \delta_k] = U^* \quad (5.24)$$

Observe that if  $U^* = 0$ , then, the above model is a  $\{1, 1, V^*, W^*\}$  DLM, so the first order polynomial constant model is embedded in the larger class of models above described.

In order to obtain valid expressions for  $U^*$ ,  $V^*$  and  $W^*$ , we analyse the process  $\eta_k$ , defined by  $\eta_k = Z_k - Z_{k-1}$ . From (5.20) and (5.21), we obtain

$$\eta_k = \epsilon_k - \epsilon_{k-1} + \delta_k$$

and, using (5.22) to (5.24):

$$Var[\eta_k] = 2V^* + 2U^* + W^* \quad (5.25)$$

$$Cov[\eta_k, \eta_{k-1}] = -V^* - U^* \quad (5.26)$$

$$Cov[\eta_k, \eta_{k-j}] = 0, \text{ for } j > 1 \quad (5.27)$$

Let  $M^* = U^* + V^*$ . We can readily see that the role played by  $V^*$  in the first model is now played by  $M^*$ , and that  $W^*$  is not changed by the introduction of  $U^*$ . Therefore, we simply get:

$$M^* = n\{V - \frac{n^2 - 1}{6}W\} \quad (5.28)$$

$$W^* = n^3W \quad (5.29)$$

that is, (5.7) and (5.8) with  $V^*$  substituted by  $M^*$ . Then, in order to find a representation for the aggregated data, we have to divide  $M^*$ , given by (5.28), in  $M^* = U^* + V^*$ , satisfying  $V^* > 0$  and  $(U^*)^2 < V^*W^*$ ,  $W^*$  being given by (5.29). From the definition of  $M^*$ , this last restriction can also be written as:

$$Q(U^*) = (U^*)^2 + W^*U^* - M^*W^* \leq 0 \quad (5.30)$$

The discriminant  $\Delta = b^2 - 4ac$  in (5.30) is

$$\Delta = n^4W\{4V + \frac{1}{3}(n^2 + 2)W\}$$

and this is always positive, as we suppose  $V, W > 0$ . Therefore, (5.30) is satisfied if we choose  $U^*$  between the roots of  $Q(\cdot)$ ; for example, we can choose the medium



point  $-W^*/2$ . Making this choice and calculating  $V^* = M^* - U^*$  from it, we arrive at:

$$U^* = -\frac{1}{2}n^3W \quad (5.31)$$

$$V^* = n\{V + \frac{2n^2 + 1}{6}W\} \quad (5.32)$$

$$W^* = n^3W \quad (5.33)$$

where we can see that restriction  $V^* > 0$  is satisfied.

It is important to observe that the result above is fundamentally different from the first one, in the sense that the representation here is by no means unique. It can be easily seen, in fact, that because of convexity of  $Q(\cdot)$ , we can choose any value of  $U^*$  between its two roots, in order to have both restrictions satisfied. In fact,  $W^*$  and  $M^*$  are the fixed parameters that do not depend on the particular representation, its values being given by (5.28) and (5.29). The variance  $W^*$  of the evolution equation is, therefore, conceptually different from  $V^*$ , in the sense that it is completely determined by the aggregation level.

We did not consider yet the problem of the initial information. Let's suppose that we are given  $(\theta_0|D_0) \sim N[m_0, C_0]$ , as in the first theorem. Of course, this will imply  $\mu_0 = E[\phi_0|D_0] = nm_0$ , as in (12). To choose  $\Gamma_0 = Var[\phi_0|D_0]$  we observe that using (5.20) to (5.24), we get

$$Var[Z_1] = \Gamma_0 + U^* + W^* + M^*$$

Substituting (5.28) and (5.29) above and comparing with (5.12), we obtain:

$$\Gamma_0 + U^* = n^2\{C_0 - \frac{n-1}{2}W\}$$

which is a variation of (5.10). Then, to guarantee positiveness of  $\Gamma_0$ , we must have

$$U^* \leq \bar{U} = n^2\{C_0 - \frac{n-1}{2}W\}$$

It is interesting to observe that we can write

$$\bar{U} = -\frac{1}{2}n^3W + n^2\{C_0 + \frac{1}{2}W\} \quad (5.34)$$

which means that (5.31) to (5.33) will always be a valid representation. Also, if  $\bar{U}$  lies between the roots of  $Q(\cdot)$ , it will be the maximum allowed value for  $U^*$ . If not, we can choose any value between the roots. It is worth observing how this links  $C_0$  to the choice of  $V^*$ . If we have a big uncertainty about  $\theta_0$ , then we can allow for a small variance  $V^*$  in the representation of the aggregated data. If uncertainty decreases (smaller  $C_0$ ), representation will be valid only if we allow for a reasonably large  $V^*$ .

In order to complete the study of this very particular case, we need to show sufficiency. This will lead us to an extension of the first theorem, as follows:

**THEOREM 2 (DLM REPRESENTATION FOR AGGREGATED DATA  $\{1, 1, V, W\}$  CASE).** Suppose we are given a process  $\{y_t\}$ , evolving in time according to the univariate  $\{1, 1, V, W\}$  model with initial information  $(\theta_0|D_0) \sim N[m_0, C_0]$ , where  $\theta_t$  is the state parameter. Consider the new series  $\{Z_k\}$ , obtained from the first by aggregating each  $n$  observations, as in (5.1). Define:

$$M^* = n\{V - \frac{n^2 - 1}{6}W\} \quad (5.35)$$

$$W^* = n^3W \quad (5.36)$$

$$Q(x) = x^2 + W^*x - M^*W^* \quad (5.37)$$

$$\bar{U} = n^2\{C_0 - \frac{n-1}{2}W\} \quad (5.38)$$

Let  $U_{10} < 0$  and  $U_{20}$  be, respectively, the smallest and biggest root in (5.37) and consider  $U_L = \min\{\bar{U}, U_{20}\} > U_{10}$ . Choose  $U^*$  such that  $U_{10} \leq U^* \leq U_L$  and define:

$$V^* = M^* - U^* > 0$$

$$\Gamma_0 = \bar{U} - U^* > 0$$

Then, the new aggregated series can be represented by the DLM

$$Z_k = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} \quad (5.39)$$

$$\begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_{k-1} \\ \epsilon_{k-1} \end{pmatrix} + \begin{pmatrix} \delta_k \\ \epsilon_k \end{pmatrix}$$

where

$$Var[\epsilon_k] = V^*$$

$$Var[\delta_k] = W^*$$

$$Cov[\epsilon_k, \delta_k] = U^*$$

and initial information

$$(\phi_0|D_0) \sim N[nm_0, \Gamma_0] \quad (5.40)$$

Moreover, if  $M^* > 0$  and  $\bar{U} > 0$ , we can simply choose  $U^* = 0$  and obtain a simpler representation.

**PROOF:** It suffices to verify that expressions for the quantities  $E[Z_j|D_0]$ ,  $Var[Z_j|D_0]$  and  $Cov[Z_j, Z_k|D_0]$  obtained from this equivalent model coincide respectively with (5.11), (5.14) and (5.18). We continue to use that same simplified notation we had in the proof of Theorem 1.

From (5.40), (5.11) is trivially verified. To arrive at (5.14), we observe that

$$\begin{aligned}
Var[Z_k] &= \Gamma_0 + kW^* + V^* + 2Cov[\phi_k, \epsilon_k] \\
&= \Gamma_0 + kW^* + V^* + 2U^* \\
&= \bar{U} + kW^* + M^*
\end{aligned} \tag{5.41}$$

Substituting (5.35), (5.36) and (5.38) in (5.41) we arrive at (5.14).

To obtain (5.18) we simply write

$$Cov[Z_k, Z_{k+j}] = \Gamma_0 + kW^* + U^* = \bar{U} + kW^* \tag{5.42}$$

and substitute (5.36) and (5.38) in (5.42).

This proves the theorem.

It is interesting to observe that  $M^* > 0$  and  $\bar{U} > 0$  is a set of conditions which is equivalent to (5.4). If both of them are satisfied, we can reduce things in order to have the result of Theorem 1. Also, we could ask about the minimum squared correlation we can have between the two noises when one of the conditions is not satisfied. By a simple study of the function  $f(x) = x^2/(a-x)$  we can conclude that we must choose  $\tilde{U} = \min\{\bar{U}, 2M^*\}$  as the covariance between  $\epsilon_k$  and  $\delta_k$  to have this minimum squared correlation. Now, it is interesting to note that for a very high aggregation level  $n$ , we will have  $M^* \approx -W^*/6$  and  $\bar{U} \approx -W^*/2$ . Thus, for a very high level of aggregation we will have approximately 0.75 as the minimum possible squared correlation between these errors. That means, for a very high level of aggregation, we can only represent the aggregated series by that same structure, if we allow for a very high correlation between the errors.

We can look to equation (5.5) as a decomposition of  $Z_k$  in a systematic component  $\phi_k$  and a random component  $\epsilon_k$ . Also, in our representation, we have  $Cov[\phi_k, \epsilon_k] = U^*$ . Therefore, an intuitive idea is that aggregation brings up a secondary effect, in the sense that the same type of systematic variation is not anymore sufficient to explain the aggregated data, after a certain level. We have, now, to introduce a correlation between the parameter  $\phi_k$  and the error  $\epsilon_k$  in order to account for the new effect.

Also observe that  $Cov[Z_k, \epsilon_k] = M^*$ . This fact helps us to understand that if  $M^* < 0$ , we have to introduce a negative correlation between  $\epsilon_k$  and  $\phi_k$ . From (5.39), it is evident that a negative correlation between  $Z_k$  and  $\epsilon_k$  can only be possible if  $\phi_k$  and  $\epsilon_k$  are negatively correlated. Therefore, in this case, we cannot any more suppose independence.

It is worth observing that representation (5.31) to (5.33) is also valid if (5.2) and (5.3) describe a *multivariate* process,  $V$  and  $W$  being now the covariance matrices of the respective errors. First, we observe that  $W^*$  as defined by (5.33) will be positive definite and that using (5.31) to (5.33) we get

$$V^* - (U^*)'(W^*)^{-1}(U^*) = n\{V + \frac{n^2 + 2}{12}W\}$$

which is also positive definite. Positive definiteness of these two matrices constitutes a necessary and sufficient condition to have a variance-covariance structure defined by  $U^*$ ,  $V^*$  and  $W^*$ . We can readily see, as well, that (5.11), (5.14) and (5.18) are still valid for the multivariate case. The expressions are obtained in exactly the same way as for the univariate process, step by step. Also, remembering (5.34), we can define the initial covariance matrix for the state space vector as being

$$\Gamma_0 = n^2\{C_0 + \frac{1}{2}W\}$$

since it is positive definite. Now (5.41) and (5.42) can be derived in the same way, and we conclude that the same representation is also valid for the multivariate process. The conditions for having a zero covariance  $U^*$  is now that  $M^*$  and  $\bar{U}$  as defined by (5.35) and (5.38) must now be positive definite matrices. For non-singular  $W$  these conditions mean all eigenvalues of  $VW^{-1}$  are greater than  $(n^2 - 1)/6$  and all eigenvalues of  $C_0W^{-1}$  are greater than  $(n - 1)/2$ . Then, it is easily seen that, as it happens in the univariate case, the simple representation with independent errors can be maintained only up to a certain aggregation level. After this level we need to introduce a covariance structure between the errors to allow for a similar representation.

### 5.3 The $\{F, I, V, W\}$ case

The next simplest case in our study is the  $\{F, I, V, W\}$  model, defined by the observational and system equations below:

$$\begin{cases} y_t = F'\theta_t + v_t & (5.43) \end{cases}$$

$$\begin{cases} \theta_t = \theta_{t-1} + \omega_t, & (5.44) \end{cases}$$

where  $F$  is a known vector and  $\theta_t$  is a commensurate parameter vector.

We first try to find a similar representation for the process  $Z_k$  defined by (5.1) from the univariate process  $y_t$  above. We state and proof the following result:

**THEOREM 3 (DLM REPRESENTATION FOR AGGREGATED DATA  $\{F, I, V, W\}$  UNIVARIATE CASE).** Suppose we are given a univariate process  $\{y_t\}$ , evolving in time

according to the  $\{F, I, V, W\}$  model with initial information  $(\theta_0|D_0) \sim N[m_0, C_0]$ , where  $\theta_t$  is the state parameter. Consider the new series  $\{Z_k\}$ , obtained from the first by aggregating each  $n$  observations, as in (5.1). Then, the new aggregated series can be represented by the DLM

$$Z_k = \begin{pmatrix} F' & 1 \end{pmatrix} \begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} \quad (5.45)$$

$$\begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_{k-1} \\ \epsilon_{k-1} \end{pmatrix} + \begin{pmatrix} \delta_k \\ \epsilon_k \end{pmatrix} \quad (5.46)$$

with initial information

$$(\phi_0|D_0) \sim N[nm_0, \Gamma_0]$$

where

$$Var[\epsilon_k] = V^* = n\{V + \frac{2n^2 + 1}{6}F'WF\}$$

$$Var[\delta_k] = W^* = n^3W$$

$$Cov[\epsilon_k, \delta_k] = U^* = -\frac{1}{2}n^3WF$$

$$Var[\phi_0|D_0] = \Gamma_0 = n^2\{C_0 + \frac{1}{2}W\}$$

Moreover, if  $M^* = V^* + F'U^* > 0$  and  $\bar{U}$  as defined by (5.38) is positive definite, we can use the simpler representation, given by:

$$Var[\epsilon_k] = V^* = n\{V - \frac{n^2 - 1}{6}F'WF\} \quad (5.47)$$

$$Var[\delta_k] = W^* = n^3W \quad (5.48)$$

$$Cov[\epsilon_k, \delta_k] = U^* = 0 \quad (5.49)$$

$$Var[\phi_0|D_0] = \Gamma_0 = n^2\{C_0 - \frac{n-1}{2}W\} \quad (5.50)$$

PROOF: We follow exactly the same lines of the anterior demonstrations. First, we observe that (5.47) to (5.49) represents a valid covariance structure. This is true since  $W^*$  defined by (5.48) is positive definite and

$$V^* - (U^*)'(W^*)^{-1}(U^*) = n\{V + \frac{n^2 + 2}{12}F'WF\}$$

is positive. Now note that  $Z_k$  can be expressed as:

$$Z_k = nF'\theta_{(k-1)n} + \sum_{i=1}^n (n+1-i)F'\omega_i + \sum_{i=1}^n v_i \quad (5.51)$$

Therefore

$$Var[Z_k] = n^2F'Var[\theta_{(k-1)n}]F + nV + \frac{n(n+1)(2n+1)}{6}F'WF$$

Now (5.44) is essentially the same as (5.3), so we can still apply (5.13) to the above expression to get:

$$Var[Z_k] = n^2 F' C_0 F + n^3 k F' W F + nV - \frac{4n^3 - 3n^2 - n}{6} F' W F \quad (5.52)$$

From (5.45) and (5.46), we have

$$Var[Z_k] = F'(\Gamma_0 + kW^*)F + V^* + F'U^* + (U^*)'F \quad (5.53)$$

Substituting (5.47) to (5.50) in (5.52) we get (5.53).

As before, we now show that the covariance between  $Z_k$  and  $Z_{k+j}$ , is given by the same expression, using both forms, for all  $k, j$ . We have

$$Z_{k+j} = nF'\theta_{(k+j-1)n} + \sum_{i=1}^n (n+1-i)F'\omega_{(k+j-1)n+i} + \sum_{i=1}^n v_{(k+j-1)n+i}$$

From (5.51) and above we have  $Cov[Z_k, Z_{k+j}] = nCov[Z_k, \theta_{(k+j-1)n}]F$ . Now we use (5.44) to obtain:

$$\begin{aligned} Cov[Z_k, Z_{k+j}] &= n^2 F' Var[\theta_{(k-1)n}]F + \frac{n^2(n+1)}{2} F' W F \\ &= n^2 F' C_0 F + n^3(k-1)F' W F + \frac{n^2(n+1)}{2} F' W F \end{aligned} \quad (5.54)$$

Also, from representation forms (5.45) and (5.46), we obtain:

$$\begin{aligned} Cov[Z_k, Z_{k+j}] &= F' Var[\phi_k]F + Cov[\epsilon_k, \theta_k]F \\ &= F'(\Gamma_0 + kW^*)F + (U^*)'F \end{aligned} \quad (5.55)$$

Using (5.48) to (5.50) in (5.55) we will get (5.54).

Suppose, now, that we have positiveness of the two quantities cited in the theorem. We consider the following representation:

$$\begin{cases} Z_k = F'\phi_k + \epsilon_k \\ \phi_k = \phi_{k-1} + \delta_k, \end{cases}$$

where we use expressions (5.47) to (5.50) for the relevant parameters. We have:

$$\begin{aligned} Var[Z_k] &= F'(\Gamma_0 + kW^*)F + V^* \\ Cov[Z_k, Z_{k+j}] &= F'(\Gamma_0 + kW^*)F \end{aligned}$$

Substituting (5.47) to (5.50) in the two expressions above, we arrive at (5.53) and (5.54).

As we would expect, the important quantity for our analysis in the last case is not  $W$  itself, but  $F'WF$ . This is due to the fact that the important parameter here, i.e., the one from which inferences can be made, is  $F'\theta_t$ , rather than  $\theta_t$  itself. Note that defining  $\kappa_t = F'\theta_t$  and  $\psi_t = F'\omega_t$  we can reduce (5.43) and (5.44) to the form:

$$\begin{cases} y_t = \kappa_t + v_t \\ \kappa_t = \kappa_{t-1} + \psi_t, \end{cases}$$

and we are back to the previous cases.

## 5.4 The $\{1, \lambda, V, W\}$ case

### 5.4.1 Basic ideas

We would like to obtain a general result for the  $\{F, G, V, W\}$  process. In other words, suppose  $y_t$  evolves as the  $\{F, G, V, W\}$  DLM below:

$$\begin{cases} y_t = F'\theta_t + v_t & (5.56) \\ \theta_t = G\theta_{t-1} + \omega_t, & (5.57) \end{cases}$$

with  $\text{Var}[v_t] = V$  and  $\text{Var}[\omega_t] = W$ .

We would like to know what kind of DLM representation can be found for the aggregated process defined by (5.1). Now, assume that this representation is given by a  $\{F^*, G^*, V^*, W^*\}$  model. It seems reasonable that we should try, in principle,  $F^* = F$  and  $G^* = G^n$ . The first choice looks reasonable, from the fact that  $F$  does not change with  $t$ , and, therefore, can be put in evidence when data is aggregated. In fact, suppose we have a representation with  $F^* = F_1$ . If we can take a diagonal nonsingular transformation  $T$  such that  $F_1 = TF$ , then, changing the state space vector using  $T$  as the reparametrization matrix will give us another representation, this time with  $F^* = F$ , and not altering the system matrix. The choice of  $G^* = G^n$  looks sensible, since from (5.56) and (5.57) we can see that the factors defining  $Z_{k+1}$  present a lag difference of  $n$  steps from their respective factors defining  $Z_k$ . For example, if we have  $W = 0$ , we can easily see that  $Z_2 = F'(\theta_{n+1} + \dots + \theta_{2n}) + \epsilon_2 = F'G^n(\theta_1 + \dots + \theta_n) + \epsilon_2$ . Intuitively, as we are aggregating  $n$  periods of time, the new transition now must correspond to  $n$  steps of the previous transition. Observe, as well, that  $Z_k$  defined by (5.1) will be such that  $E[Z_k] = G^n E[Z_{k-1}]$ . Hence, the new system matrix  $G^*$  in our representation must be the  $n$ -th power of the old system matrix,  $G$ .

#### 5.4.2 The $\{1, \lambda^n, V^*, W^*\}$ representation

We will study here the simple univariate  $\{1, \lambda, V, W\}$  case. Consider, therefore, the following DLM model for univariate  $y_t$ .

$$\begin{cases} y_t = \theta_t + v_t & (5.58) \\ \theta_t = \lambda \theta_{t-1} + \omega_t, & (5.59) \end{cases}$$

where  $Var[v_t] = V$ ,  $Var[\omega_t] = W$  and we suppose  $\lambda \neq 1$ . We try, then, to describe  $Z_k$  defined by (5.1) using the DLM below:

$$\begin{cases} Z_k = \phi_k + \epsilon_k & (5.60) \\ \phi_k = \lambda^n \phi_{k-1} + \delta_k, & (5.61) \end{cases}$$

where  $Var[\epsilon_k] = V^*$ ,  $Var[\delta_k] = W^*$  and  $(\phi_0 | D_0) \sim N[\mu_0, \Gamma_0]$ .

The basic idea is the same used for  $\lambda = 1$ . We first use (5.60) and (5.61) to get

$$\eta_k = Z_k - \lambda^n Z_{k-1} = \epsilon_k - \lambda^n \epsilon_{k-1} + \delta_k \quad (5.62)$$

Hence,  $\eta_k$  is a zero mean stationary process with  $\gamma_\eta(j) = E[\eta_k \eta_{k-j}]$  given by

$$\gamma_\eta(0) = (1 + \lambda^{2n})V^* + W^* \quad (5.63)$$

$$\gamma_\eta(1) = -\lambda^n V^* \quad (5.64)$$

$$\gamma_\eta(j) = 0 \quad \text{all } j > 1 \quad (5.65)$$

We now calculate the same autocovariances for  $\eta_k$  departing from (5.58) and (5.59). For simplicity of notation, let's consider  $k = 2$ . We have

$$\eta_2 = Z_2 - \lambda^n Z_1 = \sum_{j=1}^n (y_{n+j} - \lambda^n y_j) \quad (5.66)$$

We substitute

$$y_{n+j} - \lambda^n y_j = v_{n+j} - \lambda^n v_j + \sum_{i=j+1}^{n+j} \lambda^{n+j-i} \omega_i \quad (5.67)$$

From (5.66) and (5.67) we have

$$\eta_2 = Z_2 - \lambda^n Z_1 = \sum_{j=1}^n (v_{n+j} - \lambda^n v_j) + \sum_{j=2}^{n+1} \sum_{i=j}^{n+j-1} \lambda^{n+j-1-i} \omega_i \quad (5.68)$$



We interchange the order of summation in (5.68) to get

$$\begin{aligned}
& \sum_{j=2}^{n+1} \left( \sum_{i=j}^{n+j-1} \lambda^{n+j-1-i} \omega_i \right) = \\
& \sum_{i=2}^{n+1} \left( \sum_{j=2}^i \lambda^{n+j-1-i} \omega_i \right) + \sum_{i=n+2}^{2n} \left( \sum_{j=i+1-n}^{n+1} \lambda^{n+j-1-i} \omega_i \right) = \\
& \sum_{i=2}^{n+1} \omega_i \left( \sum_{j=2}^i \lambda^{n+j-1-i} \right) + \sum_{i=n+2}^{2n} \omega_i \left( \sum_{j=i+1-n}^{n+1} \lambda^{n+j-1-i} \right) = \\
& \sum_{i=1}^n A_i \omega_{i+1} + \sum_{i=1}^{n-1} B_i \omega_{n+i+1}
\end{aligned} \tag{5.69}$$

where

$$A_i = \sum_{j=2}^{i+1} \lambda^{n+j-2-i} = \frac{\lambda^{n-i} - \lambda^n}{1 - \lambda} \tag{5.70}$$

$$B_i = \sum_{j=i+2}^{n+1} \lambda^{j-2-i} = \frac{1 - \lambda^{n-i}}{1 - \lambda} \tag{5.71}$$

From the expressions above we can calculate the variance of  $\eta_2$ , which is also the variance of  $\eta_k$ .

$$\begin{aligned}
\gamma_\eta(0) &= (1 + \lambda^{2n})nV + \left\{ \sum_{i=1}^n A_i^2 + \sum_{i=1}^{n-1} B_i^2 \right\} W \\
&= (1 + \lambda^{2n})nV + \frac{W}{(1 - \lambda)^2} \left\{ \sum_{i=0}^{n-1} (\lambda^i - \lambda^n)^2 + \sum_{i=1}^{n-1} (1 - \lambda^i)^2 \right\} \\
&= (1 + \lambda^{2n})nV + \frac{W}{(1 - \lambda)^2} \left\{ n(1 + \lambda^{2n}) - \frac{2\lambda(1 - \lambda^{2n})}{1 - \lambda^2} \right\}
\end{aligned} \tag{5.72}$$

Now, to obtain  $\gamma_\eta(1)$  we cross the coefficients of the common lags of  $v$  and  $\omega$  in the expressions, for example, of  $\eta_2$  and  $\eta_3$ . We use (5.69) to (5.71) to rewrite (5.68) as

$$\eta_2 = -\lambda^n \sum_{j=1}^n v_j + \sum_{j=1}^n v_{n+j} + \sum_{i=1}^n A_i \omega_{i+1} + \sum_{i=1}^{n-1} B_i \omega_{n+i+1}$$

Similarly, we have

$$\eta_3 = -\lambda^n \sum_{j=1}^n v_{n+j} + \sum_{j=1}^n v_{2n+j} + \sum_{i=1}^n A_i \omega_{n+i+1} + \sum_{i=1}^{n-1} B_i \omega_{2n+i+1} \tag{5.73}$$

Therefore:

$$\begin{aligned}
\gamma_\eta(1) &= -n\lambda^n V + W \sum_{i=1}^{n-1} A_i B_i \\
&= -n\lambda^n V + \frac{W}{(1 - \lambda)^2} \left\{ \frac{\lambda(1 - \lambda^{2n})}{1 - \lambda^2} - n\lambda^n \right\}
\end{aligned} \tag{5.74}$$

If we write the expression for  $\eta_4$  based on that for  $\eta_2$  (as we did to obtain (5.73)), then, it will be clear that (5.65) holds.

Now, we are in position to obtain the values of  $V^*$  and  $W^*$ . For example, we can equate (5.74) to (5.64) to obtain  $V^*$ ; substituting its expression in (5.63) and equating this to (5.72) gives us  $W^*$ . We will first rewrite (5.72) and (5.74) in a more compact form, in order to simplify calculations. Define:

$$\begin{aligned} S_1 &= \sum_{i=1}^{n-1} (\lambda^i - \lambda^n)^2 \\ S_2 &= \sum_{i=1}^{n-1} (1 - \lambda^i)^2 \\ P &= \sum_{i=1}^{n-1} (1 - \lambda^i)(\lambda^i - \lambda^n) \\ \tilde{W} &= \frac{W}{(1 - \lambda)^2} \end{aligned}$$

Then, we can write

$$\gamma_\eta(0) = (1 + \lambda^{2n})nV + \tilde{W} \{(1 - \lambda^n)^2 + S_1 + S_2\} \quad (5.75)$$

$$\gamma_\eta(1) = -n\lambda^n V + \tilde{W}P \quad (5.76)$$

Clearly,  $S_1 + S_2 + 2P = (n - 1)(1 - \lambda^n)^2$ . Therefore, we have

$$\gamma_\eta(0) + 2\gamma_\eta(1) = n(1 - \lambda^n)^2(V + \tilde{W}) \quad (5.77)$$

But, from (5.63) and (5.64)

$$\gamma_\eta(0) + 2\gamma_\eta(1) = (1 - \lambda^n)^2 V^* + W^* \quad (5.78)$$

We compare (5.64) with (5.76) to obtain:

$$V^* = nV - \frac{P\tilde{W}}{\lambda^n} = nV - \frac{W}{(1 - \lambda)^2} \left\{ \frac{(1 - \lambda^{2n})}{\lambda^{n-1}(1 - \lambda^2)} - n \right\} \quad (5.79)$$

Let again  $r = V/W$ . Then, a condition to have this representation is

$$1 + r(1 - \lambda)^2 \geq \frac{1 - \lambda^{2n}}{n\lambda^{n-1}(1 - \lambda^2)}$$

since  $V^*$  cannot be negative.

Substituting (5.79) in (5.78) and comparing it to (5.77), we get

$$W^* = \left( \frac{1 - \lambda^n}{1 - \lambda} \right)^3 \left( \frac{1 + \lambda^n}{1 + \lambda} \right) \frac{W}{\lambda^{n-1}} \quad (5.80)$$

We readily see, from (5.80), that  $n$  must be odd when  $\lambda < 0$ , otherwise  $W^*$  will not be positive, and, therefore, the representation will not be possible. This looks reasonable if we note that, when  $n$  is even, (5.60) to (5.65) must be exactly the same, regardless of the sign of  $\lambda$ . But, the underlying process, defined by (5.58) and (5.59) is fundamentally different, depending on the sign of  $\lambda$ . Therefore, it looks reasonable that these two processes (with  $\lambda$  and with  $-\lambda$ ) cannot yield the same representation for the aggregated data. The representation will only be possible, then, for the process with  $\lambda > 0$ .

We now tackle the problem of the initial information. Suppose we are given  $(\theta_0|D_0) \sim N[m_0, C_0]$ . Then, we can obtain the distribution of  $(Z_1|D_0)$ . Observe, from (5.59) that

$$\theta_j = \lambda^j \theta_0 + \sum_{i=1}^j \lambda^{j-i} \omega_i \quad (5.81)$$

Therefore

$$\begin{aligned} Z_1 &= \sum_{j=1}^n (\theta_j + v_j) = \sum_{j=1}^n \lambda^j \theta_0 + \sum_{j=1}^n \sum_{i=1}^j \lambda^{j-i} \omega_i + \sum_{j=1}^n v_j \\ &= \frac{\lambda - \lambda^{n+1}}{1 - \lambda} \theta_0 + \frac{1}{1 - \lambda} \sum_{i=1}^n (1 - \lambda^i) \omega_{n+1-i} + \sum_{j=1}^n v_j \end{aligned} \quad (5.82)$$

Hence, we have

$$Var[Z_1] = nV + \frac{\lambda^2(1 - \lambda^n)^2}{(1 - \lambda)^2} C_0 + \tilde{W} \sum_{i=1}^n (1 - \lambda^i)^2 \quad (5.83)$$

Now, from (5.60) and (5.61) we have

$$Var[Z_1] = \lambda^{2n} \Gamma_0 + V^* + W^* = \lambda^{2n} \Gamma_0 + \gamma_\eta(0) + \lambda^n \gamma_\eta(1) \quad (5.84)$$

where we have used (5.63) and (5.64) to get the last equality. From (5.75) and (5.76), we get

$$Var[Z_1] = \lambda^{2n} \Gamma_0 + nV + \tilde{W} \left\{ \lambda^n P + S_1 + \sum_{i=1}^n (1 - \lambda^i)^2 \right\} \quad (5.85)$$

Comparing (5.83) and (5.85) we get

$$\frac{\lambda^2(1 - \lambda^n)^2}{(1 - \lambda)^2} C_0 = \lambda^{2n} \Gamma_0 + \tilde{W}(\lambda^n P + S_1)$$

which solved for  $\Gamma_0$  will give us

$$\Gamma_0 = \frac{(1 - \lambda^n)^2}{\lambda^{2n-2}(1 - \lambda)^2} \left\{ C_0 - \frac{1 - \lambda^{n-1}}{1 - \lambda^2} W \right\} \quad (5.86)$$

Then, a second condition to have this representation is

$$\frac{C_0}{W} \geq \frac{1 - \lambda^{n-1}}{1 - \lambda^2}$$

Also, from (5.82) we readily have

$$E[Z_1] = \frac{\lambda - \lambda^{n+1}}{1 - \lambda} m_0$$

But, from (5.60) and (5.61) we have  $E[Z_1] = \lambda^n \mu_0$ . Therefore, we must have:

$$\mu_0 = \frac{1 - \lambda^n}{\lambda^{n-1}(1 - \lambda)} m_0 \quad (5.87)$$

It is important to observe that the limit of (5.79), (5.80), (5.86) and (5.87) when  $\lambda \rightarrow 1$  coincide respectively with (5.7), (5.8), (5.10) and (5.9). First, we check (5.79). Three successive applications of L'Hospital's rule show that

$$\lim_{\lambda \rightarrow 1} \frac{1 - \lambda^{2n} - n\lambda^{n-1}(1 - \lambda^2)}{\lambda^{n-1}(1 - \lambda^2)(1 - \lambda)^2} = \frac{n(n^2 - 1)}{6}$$

and we have (5.7). Now, because  $\lim_{\lambda \rightarrow 1} \frac{1 - \lambda^k}{1 - \lambda} = k$ , we can easily see that (5.80) reduces to (5.8) if  $\lambda \rightarrow 1$ . Also, rewrite (5.86) as

$$\Gamma_0 = \frac{1}{\lambda^{2n-2}} \left( \frac{1 - \lambda^n}{1 - \lambda} \right)^2 \left\{ C_0 - \left( \frac{1}{1 + \lambda} \right) \left( \frac{1 - \lambda^{n-1}}{1 - \lambda} \right) W \right\} \quad (5.88)$$

From that same property, we readily see that (5.88) reduces to (5.10), when  $\lambda \rightarrow 1$ . It is also trivial from that property that (5.87) reduces to (5.9).

For completion of the result we show sufficiency, arriving at the following extension of the first theorem.

**THEOREM 4 (DLM REPRESENTATION FOR AGGREGATED DATA  $\{1, \lambda, V, W\}$  CASE).** Suppose we are given a process  $\{y_t\}$ , evolving in time according to the univariate  $\{1, \lambda, V, W\}$  model,  $\lambda \neq 1$ , with initial information  $(\theta_0 | D_0) \sim N[m_0, C_0]$ , where  $\theta_t$  is the state parameter. Consider the new series  $\{Z_k\}$ , obtained from the first by aggregating each  $n$  observations, as in (5.1) and let  $r = V/W$ . If the conditions

$$1 + r(1 - \lambda)^2 \geq \frac{1 - \lambda^{2n}}{n\lambda^{n-1}(1 - \lambda^2)} \quad (5.89)$$

$$\frac{C_0}{W} \geq \frac{1 - \lambda^{n-1}}{1 - \lambda^2} \quad (5.90)$$

are satisfied, then the new series can be represented by the  $\{1, \lambda^n, V^*, W^*\}$  model

$$\begin{cases} Z_k = \phi_k + \epsilon_k & (5.91) \\ \phi_k = \lambda^n \phi_{k-1} + \delta_k, & (5.92) \end{cases}$$

with initial information  $(\phi_0|D_0) \sim N[\mu_0, \Gamma_0]$ , where

$$V^* = nV - \frac{W}{(1-\lambda)^2} \left\{ \frac{(1-\lambda^{2n})}{\lambda^{n-1}(1-\lambda^2)} - n \right\} \quad (5.93)$$

$$W^* = \left( \frac{1-\lambda^n}{1-\lambda} \right)^3 \left( \frac{1+\lambda^n}{1+\lambda} \right) \frac{W}{\lambda^{n-1}} \quad (5.94)$$

$$\mu_0 = \frac{1-\lambda^n}{\lambda^{n-1}(1-\lambda)} m_0 \quad (5.95)$$

$$\Gamma_0 = \frac{(1-\lambda^n)^2}{\lambda^{2n-2}(1-\lambda)^2} \left\{ C_0 - \frac{1-\lambda^{n-1}}{1-\lambda^2} W \right\} \quad (5.96)$$

**PROOF:** We again will show that the first two moments of  $Z_k$  are given by the same expression, using the original model and the representation, and that the same is valid for the covariance between two observations of the aggregated series.

First, let's show this is true for the expected value of  $Z_k$ . We begin by observing that (5.82) can be generalized to

$$Z_k = \frac{\lambda(1-\lambda^n)}{1-\lambda} \theta_{(k-1)n} + \frac{1}{1-\lambda} \sum_{i=1}^n (1-\lambda^i) \omega_{kn+1-i} + \sum_{i=1}^n v_{(k-1)n+i} \quad (5.97)$$

From this, we get

$$E[Z_k] = \frac{\lambda(1-\lambda^n)}{1-\lambda} E[\theta_{(k-1)n}]$$

Using this and (5.81) we obtain

$$E[Z_k] = \frac{\lambda^{(k-1)n+1}(1-\lambda^n)}{1-\lambda} m_0 \quad (5.98)$$

Now, from (5.91) and (5.92) we will get

$$E[Z_k] = \lambda^{kn} \mu_0 \quad (5.99)$$

and substituting (5.95) in (5.99) we have (5.98). This shows that the expression for the expected value of  $Z_k$  is the same in both cases.

Now, we turn to the variance of  $Z_k$ . From (5.97)

$$Var[Z_k] = \frac{\lambda^2(1-\lambda^n)^2}{(1-\lambda)^2} Var[\theta_{(k-1)n}] + nV + \frac{W}{(1-\lambda)^2} \sum_{i=1}^n (1-\lambda^i)^2 \quad (5.100)$$

From (5.81), we get

$$Var[\theta_{(k-1)n}] = \lambda^{2(k-1)n} C_0 + \frac{1-\lambda^{2(k-1)n}}{1-\lambda^2} W \quad (5.101)$$

Substitute this in (5.100) to obtain

$$\begin{aligned} Var[Z_k] = & \frac{(1 - \lambda^n)^2}{\lambda^{2n-2}(1 - \lambda)^2} \lambda^{2kn} C_0 + nV \\ & + \frac{n(1 - \lambda^2) - 2\lambda(1 - \lambda^n) - \lambda^{2+2(k-1)n}(1 - \lambda^n)^2}{(1 - \lambda)^3(1 + \lambda)} W \end{aligned} \quad (5.102)$$

Also, from (5.91) we get

$$Var[Z_k] = Var[\phi_k] + V^* \quad (5.103)$$

But, from (5.92) we can derive an expression similar to (5.81) for  $\phi_k$ :

$$\phi_k = \lambda^{kn} \phi_0 + \sum_{i=1}^k \lambda^{(k-i)n} \delta_i \quad (5.104)$$

Therefore

$$Var[\phi_k] = \lambda^{2kn} \Gamma_0 + \frac{1 - \lambda^{2kn}}{1 - \lambda^{2n}} W^* \quad (5.105)$$

Substituting the expression above in (5.103) and using (5.93), (5.94) and (5.96) we will also obtain (5.102).

It remains to show that  $Cov[Z_k, Z_{k+j}]$  is given by the same expression in both cases. Remembering (5.97) we can write

$$\begin{aligned} Z_{k+j} = & \frac{\lambda(1 - \lambda^n)}{1 - \lambda} \theta_{(k+j-1)n} + \frac{1}{1 - \lambda} \sum_{i=1}^n (1 - \lambda^i) \omega_{(k+j)n+1-i} \\ & + \sum_{i=1}^n v_{(k+j-1)n+i} \end{aligned} \quad (5.106)$$

From (5.97) and (5.106) we see that

$$Cov[Z_k, Z_{k+j}] = \frac{\lambda(1 - \lambda^n)}{1 - \lambda} Cov[Z_k, \theta_{(k+j-1)n}] \quad (5.107)$$

Now, we can write

$$\theta_{(k+j-1)n} = \lambda^{jn} \theta_{(k-1)n} + \sum_{i=1}^{jn} \lambda^{jn-i} \omega_{(k-1)n+i}$$

which is similar to (5.81) where we shift the origin to time  $(k-1)n$ . From (5.97) and above we have

$$\begin{aligned} Cov[Z_k, \theta_{(k+j-1)n}] = & \frac{\lambda^{jn+1}(1 - \lambda^n)}{1 - \lambda} Var[\theta_{(k-1)n}] \\ & + \frac{1}{1 - \lambda} Cov \left\{ \sum_{i=1}^n (1 - \lambda^i) \omega_{kn+1-i}, \sum_{i=1}^{jn} \lambda^{jn-i} \omega_{(k-1)n+i} \right\} \end{aligned}$$

The common lags of  $\omega$  in the two sums above run from  $(k-1)n+1$  to  $(k-1)n+n = kn$ . Multiplying corresponding coefficients and adding them up, we arrive at

$$\begin{aligned} Cov[Z_k, \theta_{(k+j-1)n}] &= \frac{\lambda^{jn+1}(1-\lambda^n)}{1-\lambda} Var[\theta_{(k-1)n}] \\ &\quad + \frac{W}{1-\lambda} \sum_{i=1}^n \lambda^{jn-i}(1-\lambda^{n+1-i}) \end{aligned}$$

Using (5.101) in the expression above, and taking the result to (5.107), we arrive at

$$\begin{aligned} Cov[Z_k, Z_{k+j}] &= \frac{\lambda^{2+(2k+j-2)n}(1-\lambda^n)^2}{(1-\lambda)^2} C_0 \\ &\quad + \frac{\lambda^{1+(j-1)n}(1-\lambda^n)^2(1-\lambda^{1+(2k-1)n})}{(1-\lambda)^3(1+\lambda)} W \quad (5.108) \end{aligned}$$

Also, from (5.91) and (5.92)

$$Cov[Z_k, Z_{k+j}] = Cov[\phi_k, \phi_{k+j}] = \lambda^{jn} Var[\phi_k]$$

Substituting (5.105) above and using (5.94) and (5.96), we will also obtain (5.108). Hence, the theorem is proved.

It is interesting to analyse conditions (5.89) and (5.90). First, we observe that the right hand side of (5.89) can be rewritten as

$$\frac{\lambda(\lambda^{-n} - \lambda^n)}{n(1-\lambda^2)} = \frac{2\lambda \sinh(n \log \lambda)}{n(\lambda^2 - 1)}$$

which is an unbounded increasing function of  $n$ . Hence, we have again that same kind of result we had in the first theorem. That is, there must be an aggregation level after which we cannot anymore represent the aggregated data by the simple model. We also observe that the expression above will increase much faster with  $n$  for small values of  $|\lambda|$ . Therefore, for small values of  $|\lambda|$ , we can maintain the same structure only for small values of  $n$ . For a very small value of  $|\lambda|$ , we will not be allowed to aggregate even two observations. This looks reasonable if we observe, from (5.59), that a very small value of  $|\lambda|$  means that we are losing information on the parameter quite quickly, and the random term  $\omega_t$  will dominate the systematic effect. If information is lost quite quickly, then, very early we will need a more sophisticated model, since, it will soon be insufficient to explain the behaviour of data. Then, we expect model sophistication to be necessary for a very small aggregation level when we have a very small  $|\lambda|$ .

The right hand side of (5.90) behaves according to  $|\lambda|$ . For  $|\lambda| > 1$  it is an unbounded increasing function of  $n$ , while for  $|\lambda| < 1$  it will be a bounded function.

Therefore, for  $|\lambda| > 1$  we must always have a maximum  $n$  for which it is possible to maintain the simple structure, independently of what  $C_0$  is. On the other hand, for  $|\lambda| < 1$ , if we start with  $C_0(1 - \lambda)^2 \geq W$ , we will have (5.90) always satisfied, whatever  $n$ . This is not unreasonable if we again observe (5.102). We verify, then, that, for  $|\lambda| < 1$ , the coefficient of  $C_0$  in the expression of the variance of  $Z_k$  is a decreasing function of  $n$ , while the coefficients of  $V$  and  $W$  are increasing functions of  $n$  (the last, because of the linear term in  $n$ ). That means, for a very high aggregation level  $n$ , the knowledge about  $C_0$  does not make too much difference for explaining the variation of  $Z_k$ , in comparison with the contributions of  $v_t$  and  $\omega_t$ . If this is the case, then, it is not surprising that we can always find a suitable initial distribution for the state space in the representation (5.91)-(5.92), if  $|\lambda| < 1$ .

### 5.4.3 Model sophistication

When conditions (5.89) and (5.90) are not valid, we are forced to search for a more sophisticated model for  $Z_k$ . For  $|\lambda| < 1$ , we use again the same idea of including a zero eigenvalue in the system matrix, thus, extending our model to

$$Z_k = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} \quad (5.109)$$

$$\begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} = \begin{pmatrix} \lambda^n & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_{k-1} \\ \epsilon_{k-1} \end{pmatrix} + \begin{pmatrix} \delta_k \\ \epsilon_k \end{pmatrix} \quad (5.110)$$

where

$$Var[\epsilon_k] = V^*$$

$$Var[\delta_k] = W^*$$

$$Cov[\epsilon_k, \delta_k] = U^*$$

We will try to obtain a suitable covariance structure for  $\epsilon$  and  $\delta$ , again by analysis of  $\eta_k$ , defined as

$$\eta_k = Z_k - \lambda^n Z_{k-1} = \epsilon_k - \lambda^n \epsilon_{k-1} + \delta_k$$

Now, (5.63) to (5.65) are slightly modified by the introduction of  $U^*$ , as

$$\gamma_\eta(0) = (1 + \lambda^{2n})V^* + W^* + 2U^*$$

$$\gamma_\eta(1) = -\lambda^n(V^* + U^*)$$

$$\gamma_\eta(j) = 0 \quad \text{all } j > 1$$

Then, with  $\gamma_\eta(0)$  given by (5.72) and  $\gamma_\eta(1)$  given by (5.74), we will try to obtain valid expression for  $U^*$ ,  $V^*$  and  $W^*$ . We also remember that we must always



have  $(U^*)^2 \leq V^*W^*$  in order to have a covariance structure defined by these three quantities.

Define  $\kappa_0 = \gamma_\eta(0)$  and  $\kappa_1 = -\gamma_\eta(1)/\lambda^n$ . Then, we are looking for a point  $(U^*, V^*, W^*) \in \mathbf{R}^3$ , lying inside the elliptic cone  $(U^*)^2 \leq V^*W^*$  and also in the straight line defining the intersection of the planes given by the equations

$$2U^* + (1 + \lambda^{2n})V^* + W^* = \kappa_0 \quad (5.111)$$

$$U^* + V^* = \kappa_1 \quad (5.112)$$

Let  $\vec{r}$  be this straight line and define  $g = 1 + \lambda^{2n}$ . Also, let's choose  $(U^*, V^*, W^*)$  as our coordinate system. We can immediately verify that  $p_0 = (\kappa_1, 0, \kappa_0 - 2\kappa_1)$  and  $p_1 = (0, \kappa_1, \kappa_0 - g\kappa_1)$  are two points of  $\vec{r}$ . Here it will be important to observe that  $|\lambda| = 1$  implies  $g = 2$ , which means that  $\vec{r}$ , in this very special case, will lie on the plane  $W^* = \kappa_0 - 2\kappa_1$ , defining  $W^*$  uniquely, independently of our choice of  $U^*$  (see theorem 2). In general, if  $|\lambda| \neq 1$ , we will expect to have  $W^*$  varying with our choice for  $U^*$ .

By considering  $p_1 - p_0$ , we verify that  $\vec{r}$  has the direction of  $d = (-1, 1, 2 - g)$ . Then it can be described as the set of all points of the form  $p_0 + \alpha d, \alpha \in \mathbf{R}$ . Hence,  $\vec{r}$  is the set of the points

$$(\kappa_1 - \alpha, \alpha, \kappa_0 - 2\kappa_1 + \alpha(2 - g)), \quad \alpha \in \mathbf{R} \quad (5.113)$$

The points of  $\vec{r}$  in the cone  $(U^*)^2 \leq V^*W^*$  are those for which

$$(\kappa_1 - \alpha)^2 \leq \alpha(\kappa_0 - 2\kappa_1 + \alpha(2 - g))$$

That is

$$\lambda^{2n}\alpha^2 - \kappa_0\alpha + \kappa_1^2 \leq 0 \quad (5.114)$$

The discriminant  $\Delta = b^2 - 4ac$  in the above quadratic inequation for  $\alpha$  is

$$\begin{aligned} \Delta &= \kappa_0^2 - 4\lambda^{2n}\kappa_1^2 = (\kappa_0 + 2\lambda^n\kappa_1)(\kappa_0 - 2\lambda^n\kappa_1) \\ &= (\gamma_\eta(0) - 2\gamma_\eta(1))(\gamma_\eta(0) + 2\gamma_\eta(1)) \end{aligned}$$

Now, recalling the first line in (5.72) and the first line in (5.74), we easily have:

$$\gamma_\eta(0) - 2\gamma_\eta(1) = (1 + \lambda^n)^2 nV + A_n^2 W + W \sum_{i=1}^{n-1} (A_i - B_i)^2 > 0 \quad (5.115)$$

$$\gamma_\eta(0) + 2\gamma_\eta(1) = (1 - \lambda^n)^2 nV + A_n^2 W + W \sum_{i=1}^{n-1} (A_i + B_i)^2 > 0 \quad (5.116)$$

From (5.115) and (5.116) we readily see that  $\Delta > 0$ , which shows that  $\vec{r}$  really pierces the cone in two distinct points. Therefore

$$I_1 = \{\alpha \in \mathbf{R} | \lambda^{2n} \alpha^2 - \kappa_0 \alpha + \kappa_1^2 < 0\}$$

is a non-empty bounded interval of  $\mathbf{R}$ . Let's consider the midpoint  $\bar{\alpha}$  of  $I_1$ , given by

$$\bar{\alpha} = \frac{\kappa_0}{2\lambda^{2n}} \quad (5.117)$$

Note that  $\bar{\alpha} > 0$ , which shows that  $I_1$  has a non-empty intersection with  $(0, \infty)$  (we need to choose a positive  $\alpha$ , since  $V^*$  has to be positive; note that our parametrization identifies  $\alpha$  with  $V^*$ ). In fact, because  $\kappa_0 > 0$  and  $\kappa_1^2 > 0$ , the roots in (5.114) must be always positive. This shows that  $I_1 \subset (0, \infty)$ , which means that we do not have to worry about positivity of  $\alpha$  in  $I_1$ . Of course, positivity of  $\alpha$  in  $I_1$  automatically implies positivity of the respective  $W^*$ , since the condition defining  $I_1$  is  $(U^*)^2 \leq V^*W^*$ .

It remains to consider the problem of the initial information. From (5.109) and (5.110), we have

$$Var[Z_1] = \lambda^{2n} \Gamma_0 + V^* + W^* + 2U^*$$

where  $\Gamma_0 = Var[\phi_0]$ .

Denote  $Var[Z_1]$  by  $Q_1$ . Then, from (5.112), we have

$$\lambda^{2n} \Gamma_0 + \kappa_1 + W^* + U^* = Q_1$$

Therefore, we must choose a value of  $\alpha$  that satisfies

$$W^* + U^* \leq Q_1 - \kappa_1$$

with  $Q_1$  given by (5.83). Substituting  $U^*$  and  $W^*$  given by (5.113) we obtain

$$\alpha \geq \frac{\kappa_0 - Q_1}{\lambda^{2n}} = \alpha_0 \quad (5.118)$$

as a necessary condition for  $\alpha$ . We will show, for  $|\lambda| < 1$ , that  $I_2 = (\alpha_0, \infty)$  has a non-empty intersection with  $I_1$ , simply by showing that  $\bar{\alpha} \in I_2$ .

Suppose  $|\lambda| < 1$ , and let's show for this case that  $\bar{\alpha} \in I_2$ . To do this, we need to show that  $\bar{\alpha} > \alpha_0$ . From (5.117) and (5.118) we have

$$\bar{\alpha} - \alpha_0 = \frac{2Q_1 - \kappa_0}{2\lambda^{2n}}$$

Therefore, we need to verify that  $2Q_1 - \kappa_0 > 0$  when  $|\lambda| < 1$ . From (5.83) and (5.75) we have

$$\begin{aligned} 2Q_1 - \kappa_0 &= 2nV + \frac{2\lambda^2(1 - \lambda^n)^2}{(1 - \lambda)^2}C_0 + 2\tilde{W} \sum_{i=1}^n (1 - \lambda^i)^2 - (1 + \lambda^{2n})nV \\ &\quad - \tilde{W} \{(1 - \lambda^n)^2 - S_1 - S_2\} \\ &= \frac{2\lambda^2(1 - \lambda^n)^2}{(1 - \lambda)^2}C_0 + (1 - \lambda^{2n})nV + \tilde{W} \{(1 - \lambda^n)^2 + S_2 - S_1\} \end{aligned}$$

We, then, have to show that the coefficient of  $\tilde{W}$  in the anterior expression is always positive for  $|\lambda| < 1$ , and we will have the required result. This coefficient is

$$\begin{aligned} (1 - \lambda^n)^2 + S_2 - S_1 &= \sum_{i=1}^n \{(1 - \lambda^i)^2 - (\lambda^i - \lambda^n)^2\} \\ &= (1 - \lambda^n) \sum_{i=1}^n (1 - 2\lambda^i + \lambda^n) \\ &= (1 - \lambda^n) \frac{(2 - n)\lambda^{n+1} + n\lambda^n - (n + 2)\lambda + n}{1 - \lambda} \end{aligned}$$

Then, all we are left with is to show that the polynomial  $\pi_n(\lambda) = (2 - n)\lambda^{n+1} + n\lambda^n - (n + 2)\lambda + n$  is always positive for  $\lambda \in (-1, 1)$ . Since  $\pi_n(0) = n > 0$ , we just need to demonstrate the following

**LEMMA.** *For any fixed  $n$ , the polynomial  $\pi_n(\lambda) = (2 - n)\lambda^{n+1} + n\lambda^n - (n + 2)\lambda + n$  has no roots in  $(-1, 1)$ .*

**PROOF:** We have  $\pi_1(\lambda) = \lambda^2 - 2\lambda + 1 = (\lambda - 1)^2$ . Therefore, for  $n = 1$ , the lemma holds (as it should, for  $n = 1$  corresponds to no aggregation at all).

Also, we have  $\pi_2(\lambda) = 2\lambda^2 - 4\lambda + 2 = 2(\lambda - 1)^2$ . Then, for  $n = 2$  the lemma holds.

Let's consider  $n > 2$  and suppose  $\pi_n(\lambda)$  has a root,  $\lambda_0$ , in  $(-1, 1)$ . We have  $\pi_n(1) = 2 - n + n - (n + 2) + n = 0$ . Therefore, by the mean value theorem, we can find  $\lambda_1 \in (\lambda_0, 1)$  which is a root of the derivative  $\pi'_n(\lambda)$ .

Let's calculate  $\pi'_n(\lambda)$ , this is

$$\pi'_n(\lambda) = (2 - n)(n + 1)\lambda^n + n^2\lambda^{n-1} - (n + 2) \quad (5.119)$$

We verify that  $\pi'_n(1) = (2 - n)(n + 1) + n^2 - (n + 2) = 0$ . Then, by the same reason, there must exist  $\lambda_2 \in (\lambda_1, 1)$  which is a root of  $\pi''_n(\lambda)$ . This is

$$\begin{aligned} \pi''_n(\lambda) &= (2 - n)(n + 1)n\lambda^{n-1} + n^2(n - 1)\lambda^{n-2} \\ &= \lambda^{n-2} \{n^2(n - 1) - (n - 2)(n + 1)n\lambda\} \end{aligned}$$

We readily see that the polynomial above has  $n - 2$  zero roots and a single root given by

$$\frac{n^2(n-1)}{(n-2)(n+1)n} = \frac{n^2-n}{n^2-n-2} > 1$$

Then, we necessarily have  $-1 < \lambda_0 < \lambda_1 < 0$ . We can also deduce that the concavity of  $\pi_n(\lambda)$  does not change in  $(-1, 0)$ , since  $\pi_n''(\lambda)$  has no roots in this interval.

Now, observe that  $\pi_n(-1) = 2 - n - n + n + 2 + n = 4 > 0$  if  $n$  is odd and  $\pi_n(-1) = n - 2 + n + n + 2 + n = 4n > 0$  if  $n$  is even. Also,  $\pi_n(0) = n > 0$ . If  $\pi_n(-1) > 0$  and  $\pi_n(0) > 0$ , then, the concavity of  $\pi_n(\lambda)$  in  $(-1, 0)$  (which, as we have seen, does not change) cannot be negative. If this was the case, then we would have a positive minimum (at  $\lambda = -1$  or  $\lambda = 0$ ) of  $\pi_n(\lambda)$  in  $[-1, 0]$ , which is not true, since we assume a root. Then,  $\pi_n(\lambda)$  must be a convex function in  $(-1, 0)$ . This will allow for a root. But, then,  $\lambda_1$  must be a local minimum of  $\pi_n(\lambda)$ . That is,  $\pi_n'(0) > 0$ . But, from (5.119), we have  $\pi_n'(0) = -(n+2) < 0$  and this is a contradiction. Therefore, we cannot have a root of  $\pi_n(\lambda)$  in  $(-1, 1)$ , and the lemma is proved for  $n > 2$ .

Now, we show sufficiency, extending Theorem 4 in the same way we extend Theorem 1 to Theorem 2.

**THEOREM 5 (DLM REPRESENTATION FOR AGGREGATED DATA  $\{1, \lambda, V, W\}$  CASE,  $|\lambda| < 1$ ).** *Suppose we are given a process  $\{y_t\}$ , evolving in time according to the univariate  $\{1, \lambda, V, W\}$  model,  $|\lambda| < 1$ , with initial information  $(\theta_0|D_0) \sim N[m_0, C_0]$ , where  $\theta_t$  is the state parameter. Consider the new series  $\{Z_k\}$ , obtained from the first by aggregating each  $n$  observations, as in (5.1). Define:*

$$\mu_0 = \frac{1 - \lambda^n}{\lambda^{n-1}(1 - \lambda)} m_0 \quad (5.120)$$

$$S_1 = \sum_{i=1}^n (\lambda^i - \lambda^n)^2 \quad (5.121)$$

$$S_2 = \sum_{i=1}^n (1 - \lambda^i)^2 \quad (5.122)$$

$$P = \sum_{i=1}^n (1 - \lambda^i)(\lambda^i - \lambda^n) \quad (5.123)$$

$$\tilde{W} = \frac{W}{(1 - \lambda)^2} \quad (5.124)$$

$$\kappa_0 = (1 + \lambda^{2n})nV + (S_1 + S_2)\tilde{W} \quad (5.125)$$

$$\kappa_1 = nV - \lambda^{-n}P\tilde{W} \quad (5.126)$$

$$Q_1 = nV + \frac{\lambda^2(1 - \lambda^n)^2}{(1 - \lambda)^2}C_0 + S_2\tilde{W} \quad (5.127)$$

$$I_1 = \{\alpha \in \mathbf{R} | \lambda^{2n}\alpha^2 - \kappa_0\alpha + \kappa_1^2 < 0\} \quad (5.128)$$

$$I_2 = \{\alpha \in \mathbf{R} | \alpha \geq \lambda^{-2n}(\kappa_0 - Q_1)\} \quad (5.129)$$

Choose  $V^* \in I^* = I_1 \cap I_2 \neq \emptyset$  and define

$$U^* = \kappa_1 - V^* \quad (5.130)$$

$$W^* = \kappa_0 - 2\kappa_1 + (1 - \lambda^{2n})V^* \quad (5.131)$$

$$\Gamma_0 = V^* - \lambda^{-2n}(\kappa_0 - Q_1) \quad (5.132)$$

Then, the new aggregated series can be represented by the DLM

$$Z_k = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} \quad (5.133)$$

$$\begin{pmatrix} \phi_k \\ \epsilon_k \end{pmatrix} = \begin{pmatrix} \lambda^n & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_{k-1} \\ \epsilon_{k-1} \end{pmatrix} + \begin{pmatrix} \delta_k \\ \epsilon_k \end{pmatrix} \quad (5.134)$$

where

$$Var[\epsilon_k] = V^*$$

$$Var[\delta_k] = W^*$$

$$Cov[\epsilon_k, \delta_k] = U^*$$

and initial information

$$(\phi_0 | D_0) \sim N[\mu_0, \Gamma_0]$$

Moreover, if  $\kappa_1 \in I^*$ , we can simply choose  $V^* = \kappa_1$  and obtain a simpler representation, where  $U^* = 0$ .

PROOF: We again have to show that the first two moments for any vector of observations of  $\{Z_k\}$  calculated from the representation we give are consistent with those obtained from the original model. We already know that this is true for  $E[Z_k]$ . The proof is exactly the same as in Theorem 4. Now, the variance of  $Z_k$  is calculated from our representation as

$$Var[Z_k] = Var[\phi_k] + V^* + 2U^* \quad (5.135)$$

where  $Var[\phi_k]$  can be obtained from (5.105). We substitute (5.131) and (5.132) in (5.105) (note that (5.131) and (5.132) are positive if we choose  $V^* \in I^*$ ). This substitution leads us to

$$Var[\phi_k] = V^* + \lambda^{2(k-1)n}(Q_1 - \kappa_0) + \frac{1 - \lambda^{2kn}}{1 - \lambda^{2n}}(\kappa_0 - 2\kappa_1)$$

We use this together with (5.130) in (5.135) to get

$$Var[Z_k] = \lambda^{2(k-1)n}(Q_1 - \kappa_0) + \frac{1 - \lambda^{2kn}}{1 - \lambda^{2n}}(\kappa_0 - 2\kappa_1) + 2\kappa_1 \quad (5.136)$$

and we see that  $Var[Z_k]$  does not depend of our choice of  $V^*$  (as expected). Now, we substitute (5.125) to (5.127), and using (5.121) to (5.124) we will get (5.102).

Finally, we calculate  $Cov[Z_k, Z_{k+j}]$  using our representation. We have

$$Cov[Z_k, Z_{k+j}] = Cov[Z_k, \phi_{k+j}]$$

We can use a slight modification of (5.104) to obtain this covariance. We have

$$\phi_{k+j} = \lambda^{jn}\phi_k + \sum_{i=1}^j \lambda^{(j-i)n}\delta_{k+i}$$

Therefore

$$Cov[Z_k, \phi_{k+j}] = \lambda^{jn}Cov[Z_k, \phi_k] = \lambda^{jn}(Var[\phi_k] + U^*)$$

Use (5.130) to get

$$Cov[Z_k, \phi_{k+j}] = \lambda^{jn}\left\{\lambda^{2(k-1)n}(Q_1 - \kappa_0) + \frac{1 - \lambda^{2kn}}{1 - \lambda^{2n}}(\kappa_0 - 2\kappa_1) + \kappa_1\right\}$$

and, again, the expression we get does not depend on  $V^*$  (as expected). Now use (5.136) to get

$$Cov[Z_k, \phi_{k+j}] = \lambda^{jn}(Var[Z_k] - \kappa_1)$$

Substituting (5.102) and (5.126), then using (5.123) and (5.124), we will get (5.108).

This proves the theorem.

It is important to note that the conditions  $\kappa_1 \in I_1$  and  $\kappa_1 \in I_2$  in Theorem 5 reduce respectively to (5.89) and (5.90) in Theorem 4. In this case, we can make  $\alpha = \kappa_1$  and have a simpler representation. First, let's verify that  $\kappa_1 \in I_1$  reduces to (5.89). If this is the case, then, we have, by definition of  $I_1$ ,

$$\lambda^{2n}\kappa_1^2 - \kappa_0\kappa_1 + \kappa_1^2 < 0$$

which can be rewritten as

$$\kappa_1\{(1 + \lambda^{2n})\kappa_1 - \kappa_0\} < 0 \quad (5.137)$$

Because  $I_1 \subset (0, \infty)$ , we will have  $\kappa_1 > 0$  and

$$\kappa_0 - (1 + \lambda^{2n})\kappa_1 > 0$$

From (5.79) and (5.126), we readily see that  $\kappa_1 > 0$  is exactly (5.89). Now, observe that (5.78) can be rewritten as

$$\kappa_0 - 2\lambda^n \kappa_1 = (1 - \lambda^n)^2 \kappa_1 + W^* \quad (5.138)$$

and (5.138) becomes

$$W^* = \kappa_0 - (1 + \lambda^{2n})\kappa_1$$

Here enters the fact that the  $W^*$  of Theorem 4 must be positive; we cannot have  $\kappa_1 \in I_1$  if  $\lambda < 0$  and  $n$  is even.

It is also easy to see that  $\kappa_1 \in I_2$  reduces to (5.90). From the definition of  $I_2$ , this means

$$Q_1 \geq \kappa_0 - \lambda^{2n} \kappa_1$$

Now observe that (5.84) can be rewritten as

$$Q_1 = \lambda^{2n} \Gamma_0 + \kappa_0 - \lambda^{2n} \kappa_1$$

and this is exactly how (5.90) was obtained.

#### 5.4.4 The importance of correlation

Theorem 5 states that if  $\kappa_1 \in I^*$  we can have a simple representation for the aggregated series, where the errors of the evolutionary and system equations are uncorrelated. On the other hand, if  $\kappa_1 \notin I^*$  we need to introduce a correlation between these errors. Therefore, a relevant question here concerns the importance of correlation in practical terms. In other words, we can ask if the variance of the one-step ahead prediction error will increase significantly if we use a model where the two errors are uncorrelated, when  $\kappa_1 \notin I^*$ .

Now, from (5.113), the squared correlation between these errors is

$$S(\alpha) = \frac{(\kappa_1 - \alpha)^2}{\alpha(\kappa_0 - 2\kappa_1 + \alpha(2 - g))} \quad (5.139)$$

So, for  $\kappa_1 < 0$  and positive values of  $\alpha$ ,  $S(\alpha)$  attains a minimum at  $\alpha = \bar{\alpha}$ , where

$$\bar{\alpha} = \frac{-\kappa_1(\kappa_0 - 2\kappa_1)}{\kappa_0 - 2\lambda^{2n}\kappa_1}$$

Substituting this in (5.139), we see that the minimum possible squared correlation between the errors is given by

$$\bar{S} = \frac{-4\kappa_1}{\kappa_0 - 2\kappa_1} - 4(1 - \lambda^{2n}) \left( \frac{\kappa_1}{\kappa_0 - 2\kappa_1} \right)^2 \quad (5.140)$$

But, from the definition of  $\kappa_0$  and  $\kappa_1$

$$\frac{\kappa_1}{\kappa_0 - 2\kappa_1} = \frac{-\gamma_\eta(1)}{\lambda^n \kappa_0 + 2\gamma_\eta(1)}$$

From (5.72) and (5.74) we see that  $\lambda^n \kappa_0$  converges to zero and  $\gamma_\eta(1)$  converges to a positive value as  $n$  increases. Therefore, for a large value of  $n$  we can write

$$\frac{\kappa_1}{\kappa_0 - 2\kappa_1} = -\frac{1}{2} + \epsilon$$

where  $\epsilon$  has a very small absolute value. Using this in (5.140), we arrive at

$$\bar{S} = 1 - 4\epsilon^2 + \lambda^{2n}(1 - 2\epsilon)^2$$

Hence, we can see that for a large value of  $n$  we must introduce a high correlation between the residuals in order to have the representation of Theorem 5.

Now we can ask if we necessarily lose too much in practical terms by not introducing this high correlation. In other words, we ask if there is a significant increase in the one step ahead prediction error. The correlation structure of this error can be obtained as follows. First, we note from (5.133) and (5.134) that

$$\eta_k = (1 - hB)Z_k = (1 - \beta^*B)a_k \quad (5.141)$$

where  $h = \lambda^n$ ,  $B$  is the backshift operator,  $|\beta^*| < 1$  and  $a_k$  is a sequence of i.i.d random variables. Suppose we try to forecast the process by using another model, different from the true one. The one step ahead prediction error  $e_k$ , given by

$$e_k = Z_k - hm_{k-1} \quad (5.142)$$

where  $m_{k-1}$  is the expected value of  $Z_k$  given all information up to  $k - 1$ , will not be a sequence of independent errors under the wrong model. We apply the operator  $(1 - hB)$  to this error to obtain

$$(1 - hB)e_k = (1 - hB)Z_k - h(1 - hB)m_{k-1}$$

From the updating equations of the normal distribution we can write

$$(1 - hB)e_k = (1 - hB)Z_k - h(1 - \beta)Be_k$$



where  $\beta$  is the discount factor defining the assumed model for  $Z_k$ . This gives us

$$(1 - hB)Z_k = (1 - h\beta B)e_k$$

Comparing (5.141) with above and using the fact that  $|h\beta| < 1$ , we arrive at

$$e_k = (1 - h\beta B)^{-1}(1 - \beta^* B)a_k$$

The autocovariance generating function of the error  $e_k$  will be given by

$$\begin{aligned}\gamma_e(B) &= \sum_{i=-\infty}^{\infty} \gamma_e(i)B^i \\ &= (1 - h\beta B)^{-1}(1 - h\beta B^{-1})^{-1}(1 - \beta^* B)(1 - \beta^* B^{-1})V(a)\end{aligned}$$

if  $V(a)$  is the variance of  $a_k$ . We rewrite this as

$$\begin{aligned}\frac{\gamma_e(B)}{V(a)} &= \left( \sum_{i=0}^{\infty} (h\beta)^i B^i \right) \left( \sum_{i=0}^{\infty} (h\beta)^i B^{-i} \right) (1 - \beta^* B)(1 - \beta^* B^{-1}) \\ &= \frac{1}{1 - h^2 \beta^2} \left( \sum_{i=-\infty}^{\infty} (h\beta)^{|i|} B^i \right) (1 - \beta^* B)(1 - \beta^* B^{-1}) \\ &= \frac{1}{1 - h^2 \beta^2} \left( \sum_{i=-\infty}^{\infty} (h\beta)^{|i|} B^i \right) (-\beta^* B^{-1} + 1 + (\beta^*)^2 - \beta^* B)\end{aligned}\tag{5.143}$$

The ratio  $R$  of the variance  $V(e)$  of  $e_k$  and  $V(a)$  will be the term in  $B^0$  in the anterior expression. This ratio can be seen as a function of  $\beta$ , and we want to choose  $\beta$  to get the smallest  $R$ . We have

$$R(\beta) = \frac{V(e)}{V(a)} = \frac{1 + (\beta^*)^2 - 2h\beta\beta^*}{1 - h^2 \beta^2}\tag{5.144}$$

Defining  $\tilde{\beta} = \beta^*/h$ , we get

$$\frac{\partial R}{\partial \beta} = \frac{2h^2(1 - h\beta^*\beta)}{(1 - h^2 \beta^2)^2}(\beta - \tilde{\beta})\tag{5.145}$$

and the derivative will have the same sign of  $\beta - \tilde{\beta}$ .

We analyse now the problem, according to the sign of  $\tilde{\beta}$ .

CASE 1  $\tilde{\beta} > 0$

Suppose  $\tilde{\beta}$  is positive. That means  $\kappa_1$  is positive, since we observe from (5.141) that  $\gamma_\eta(1) = -\beta^*V(a)$ , and, therefore,  $\kappa_1 = \tilde{\beta}V(a)$ . We divide this case into two subcases. We first analyse what happens when  $\lambda > 0$  or when  $\lambda < 0$  and  $n$  is odd. In other words, we first analyse what happens when a simple representation with uncorrelated errors can be possible for the aggregated data. We shall show that we necessarily have  $\tilde{\beta} < 1$  in this first subcase. Then, we analyse the second case, when  $\lambda < 0$  and  $n$  is even, that is, when that simple representation is never possible. We show that, for this second case,  $\tilde{\beta} > 1$ .

CASE 1.1 ( $\lambda > 0$ ) or ( $\lambda < 0$  and  $n$  is odd)

For this first subcase of  $\tilde{\beta} > 0$ , we want to show that  $\tilde{\beta} < 1$ . We begin by obtaining from (5.141) that

$$\gamma_\eta(0) = (1 + (\beta^*)^2)V(a) \quad (5.146)$$

$$\gamma_\eta(1) = -\beta^*V(a) \quad (5.147)$$

and, therefore,  $\beta^*$  is the root with absolute value less than one in the equation

$$\gamma_\eta(1)(\beta^*)^2 + \gamma_\eta(0)\beta^* + \gamma_\eta(1) = 0 \quad (5.148)$$

This gives

$$\tilde{\beta} = \frac{\kappa_0}{2\lambda^{2n}\kappa_1} - \sqrt{\left(\frac{\kappa_0}{2\lambda^{2n}\kappa_1}\right)^2 - \frac{1}{\lambda^{2n}}}$$

Therefore, an equivalent condition for  $\tilde{\beta} < 1$  will be

$$\frac{\kappa_0}{\kappa_1} > 1 + \lambda^{2n}$$

and this can be put because  $|\lambda| < 1$ . Now, remember from (5.75) and (5.76) that

$$\frac{\kappa_0}{\kappa_1} = \frac{(1 + \lambda^{2n})nV + \tilde{W}(S_1 + S_2)}{nV - \lambda^{-n}P\tilde{W}}$$

with  $S_1$ ,  $S_2$ ,  $P$  and  $\tilde{W}$  given by (5.121) to (5.124). Therefore, if we can guarantee that  $\lambda^{-n}P$  is positive, we can proceed as

$$\frac{\kappa_0}{\kappa_1} > \frac{(1 + \lambda^{2n})nV}{nV - \lambda^{-n}P\tilde{W}} > \frac{(1 + \lambda^{2n})nV}{nV} = 1 + \lambda^{2n}$$

Then, all it is remaining is to show that  $\lambda^{-n}P$  will be positive in this first subcase. For  $\lambda > 0$  this is clearly true, since we can readily see from (5.123) that  $P$  will be a sum of positive terms. When  $\lambda < 0$  and  $n$  is odd, we consider the expression

$$\lambda^{-n}P = \frac{(1 - \lambda^{2n})}{\lambda^{n-1}(1 - \lambda^2)} - n \quad (5.149)$$

and this is an even function of  $\lambda$ , since  $n$  is odd. Then, as it is positive for positive  $\lambda$ , it will also be for  $\lambda < 0$ . This shows that  $\tilde{\beta} \in (0, 1)$ .

CASE 1.2  $\lambda < 0$  and  $n$  is even

Now, we want to verify that  $\tilde{\beta} > 1$  in this second subcase. Equivalently, we want to verify that

$$\frac{\kappa_0}{\kappa_1} < 1 + \lambda^{2n}$$

and this means

$$\frac{(1 + \lambda^{2n})nV + \tilde{W}(S_1 + S_2)}{nV - \lambda^{-n}P\tilde{W}} < 1 + \lambda^{2n}$$

which can be reduced to

$$S_1 + S_2 < -(1 + \lambda^{2n})\lambda^{-n}P \quad (5.150)$$

and since  $S_1 + S_2 + 2P = n(1 - \lambda^n)^2$ , we can reduce (5.150) to

$$\lambda^{-n}P + n < 0 \quad (5.151)$$

But, from (5.149), we readily get

$$\lambda^{-n}P + n = \frac{1 - \lambda^{2n}}{\lambda^{n-1}(1 - \lambda^2)}$$

and, because  $n - 1$  is odd, (5.151) is verified.

Now, we analyse the first subcase. When  $\tilde{\beta} \in (0, 1)$ , we can use  $\beta = \tilde{\beta}$  in our simplified model. From (5.145), we see that this will be a minimum point of  $R(\beta)$ . Also, from (5.144), we will have  $R(\tilde{\beta}) = 1$ . In fact, we can expect, in this case, to have  $e_k$  and  $a_k$  with the same distribution. Therefore, although we do not rigorously have that simple representation, we can, in principle, do as good as the true model, in the sense that we may have the same one step ahead error variance. It is interesting to observe, from what was developed, that  $\tilde{\beta} \in (0, 1)$  means, in fact,  $\kappa_1 > 0$  and  $\kappa_0/\kappa_1 > 1 + \lambda^{2n}$ . Then, from (5.137), the condition  $\tilde{\beta} \in (0, 1)$  is equivalent to  $\kappa_1 \in I_1$ , in Theorem 5. Hence, what we can conclude here is that the restriction  $\kappa_1 \in I_1$  in Theorem 5 is structurally much more important than the other one, namely,  $\kappa_1 \in I_2$ . If  $\kappa_1$  is in  $I_1$ , but not in  $I_2$ , we do not have the representation in terms of uncorrelated errors, but we can use a model with uncorrelated errors that will have the same forecasting efficiency as the true one.

#### CASE 2 $\tilde{\beta} < 0$

When  $\tilde{\beta} < 0$ , then (5.145) will always be positive in  $(0, 1)$ . That means our best choice will be  $\beta = 0$  for the simpler model, and that is equivalent to treating the aggregated series  $Z_k$  as a process of the form

$$Z_k = \lambda^n Z_{k-1} + \delta_k \quad (5.152)$$

where  $\delta_k$  is a sequence of uncorrelated errors. Observe that the right hand side of (5.89) is an increasing function of  $n$ , for fixed  $\lambda$ . That means, when increasing the aggregation level we must arrive at a certain point after which the best thing we

can do, if we want to maintain the simple model, is to treat the aggregated data as a simple random walk, as in (5.152). This result seems reasonable if we think in terms of forecasting. Observe that  $E[Z_k|D_{k-1}] = \lambda^n Z_{k-1}$  under (5.152), which means that the last information is everything we use to forecast  $Z_k$ . Increasing the aggregation level means increasing the length of time over which data is aggregated. In accordance with intuition, we expect to arrive at a point where this interval of time gets sufficiently wide, so that the information obtained in the last interval should be sufficient for a good forecast for the next interval, the information from the preceding intervals becoming totally irrelevant.

From (5.144),  $R(0) = 1 + (\beta^*)^2$ , which means the behaviour of  $\beta^*$  will determine how much we are losing by not considering the correlation when trying to forecast the aggregated data. Let's first study the case where  $\lambda = 1$ . We can see, for this case, that  $\beta^*$  will be a decreasing function of  $n$ . Remember from (5.25) and (5.26) that equation (5.148) will become

$$M^*(\beta^*)^2 - (2M^* + W^*)\beta^* + M^* = 0 \quad (5.153)$$

The product of the roots of the above equation is always one, the sum of the roots being  $2 + W^*/M^*$ . Let again  $r = V/W$ . Then, from (5.28) and (5.29), the sum of the roots will be given by:

$$2 + 6 \left\{ \frac{6r + 1}{n^2} - 1 \right\}^{-1}$$

From the above expression we can see what happens to the sum of roots, and, therefore, to  $\beta^*$ . While restriction (5.4) is not broken, the term in brackets is positive, and the sum of the roots will be a positive increasing function of  $n$ . This sum is  $\beta^* + (\beta^*)^{-1}$ . Then,  $\beta^*$  must be a positive decreasing function of  $n$ . After the restriction is broken (and this is the case we are interested in) the term in brackets becomes negative and it decreases (since it increases in absolute value to  $-1$ ). Therefore the sum of the roots increases to  $-4$ . Because  $\beta^*$  is the root with absolute value less than one, it must be a decreasing function of  $n$ .

We have proved that when  $\lambda = 1$  and  $\beta^*$  is negative, we have  $R(0) = 1 + (\beta^*)^2$  as an increasing function of  $n$ . That means, when  $\lambda = 1$  we are never worse than the limit case. Let's calculate this limit case. In the limit the sum of the roots of (5.153) is equal to  $-4$ . Because the product of the roots is one, we have the limit  $\beta_L^*$  as a root of the equation

$$(\beta_L^*)^2 + 4\beta_L^* + 1 = 0 \quad (5.154)$$

and this gives  $\beta_L^* = \sqrt{3} - 2$ . Now, from (5.154), we have  $1 + (\beta_L^*)^2 = -4\beta_L^* = 4(2 - \sqrt{3})$ . This is approximately 1.072, meaning that the simple model defined by putting  $\beta = 0$  is explaining a large fraction of the total variation. Therefore, we don't lose too much in terms of forecasting by using the simple model, although the exact model has to consider a reasonably large correlation between the errors.

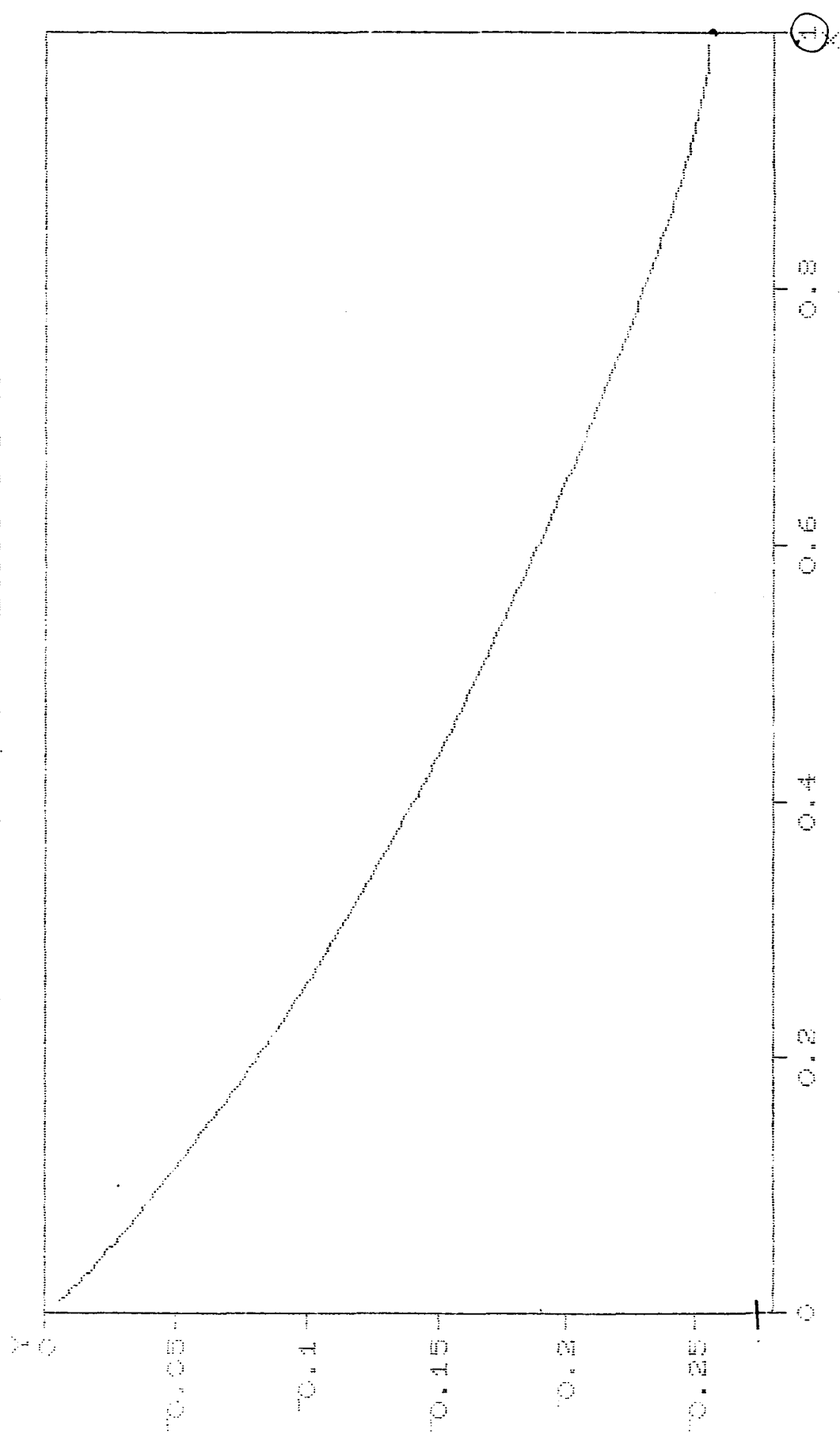
The general case (where  $\lambda \neq 1$ ) is more complex. We have already seen that  $\gamma_\eta(0)$  goes to infinity and  $\gamma_\eta(1)$  converges to a finite value as  $n$  increases. Therefore, the sum of the roots in (5.148) must diverge with  $n$ , and, consequently  $\beta^*$  must converge to zero. Let's think of the case where  $\lambda$  is positive. We will certainly have  $\beta^* > 0$  for  $n = 1$ . As  $n$  increases, (5.89) will be eventually broken, and  $\beta^*$  will become negative. But, if  $\beta^*$  converges to zero, we necessarily must have a negative minimum value  $\beta_m^*$  of  $\beta^*$  at a certain  $n_0(\lambda)$  for fixed  $\lambda > 0$ . If it can be proved that  $\beta_m^*$  is a decreasing function of  $\lambda$  in  $(0, 1)$ , then we can guarantee that, for  $\lambda > 0$ , we are never worse than that limit  $\beta_L^* = \sqrt{3} - 2$ . The next three pages show the graphs (G 5.1), (G 5.2) and (G 5.3) of  $\beta_m^*$  against  $\lambda$ , plotted for some set of fixed values of  $V$  and  $W$ . Each graph is obtained calculating the value of  $\beta_m^*$  as a function of  $\lambda$  at 201 equally spaced points with  $\lambda_1 = 0.01$  and  $\lambda_{201} = 0.99$ . From the graphs, it looks apparent that  $\beta_m^*$  must be a monotonic decreasing function of  $\lambda$ , whenever it is negative. The conclusion is, then, that, for  $\lambda > 0$ , we can always have a good performance by using the simple model with uncorrelated errors when we try to forecast the aggregated data, explaining at least  $(8 - 4\sqrt{3})^{-1} \approx 0.933$  of total variation.

Before proceeding to the next topic, we must mention here the first order autocorrelation of the one step ahead forecasting error  $e_k$ . Calculating the term in  $B^1$  in (5.143) and using (5.144) we get the first order autocorrelation  $\rho_e(1)$  as

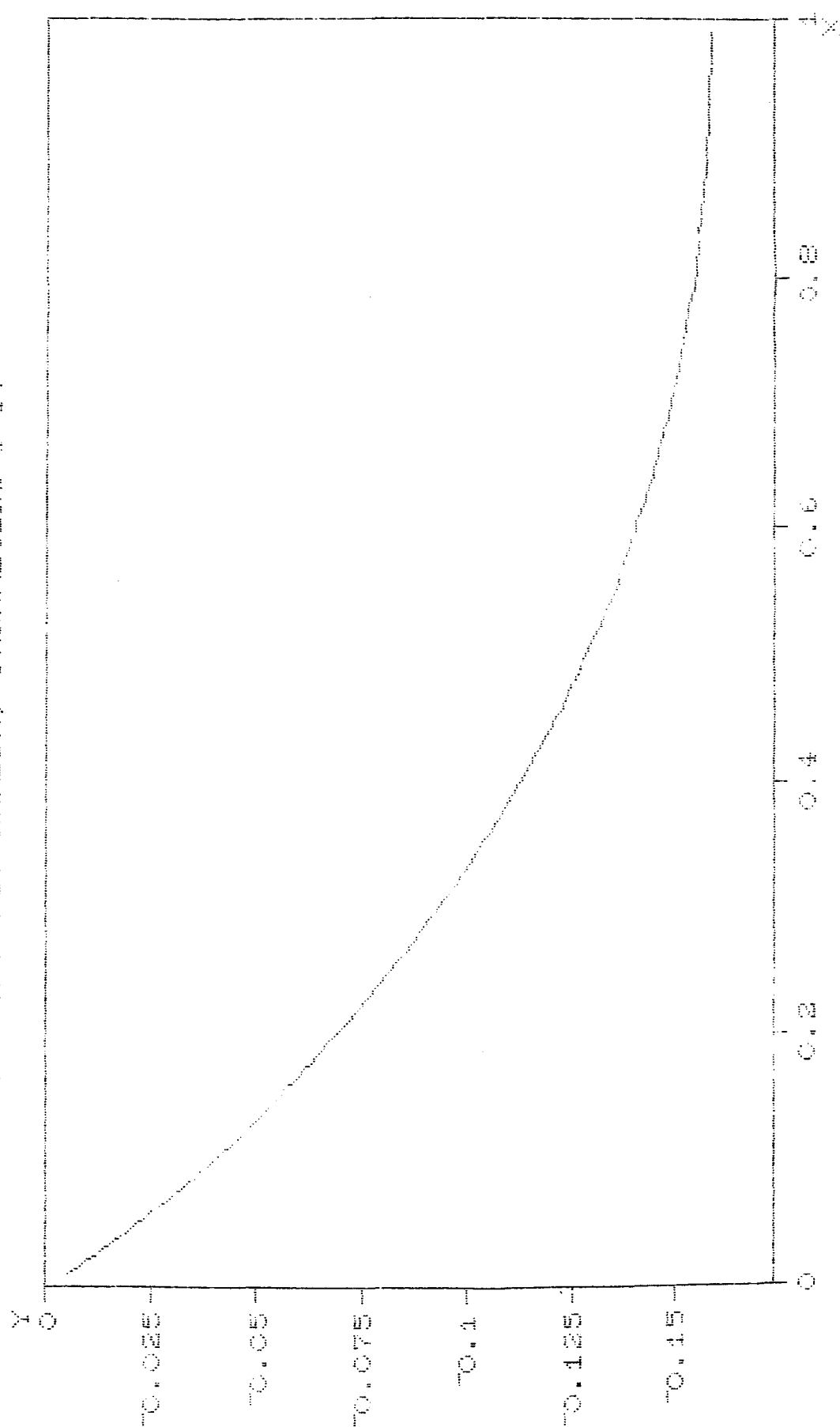
$$\rho_e(1) = \frac{(1 - h\beta\beta^*)(h\beta - \beta^*)}{1 + (\beta^*)^2 - 2h\beta\beta^*}$$

If we can make  $\beta = \tilde{\beta}$ , then, it is clear from the above expression that we will get  $\rho_e(1) = 0$ , as expected. If not, we will choose  $\beta = 0$ , which will give  $\rho_e(1) = -\beta^*/(1 + (\beta^*)^2)$ . This, from (5.146) and (5.147), is just  $\rho_\eta(1)$ , as we can expect from (5.152). Observe that this correlation goes to zero as  $n$  increases, which looks quite reasonable; if we aggregate for a larger interval of time, we expect the simple model that just considers the last observation to forecast the next to be more sensible.

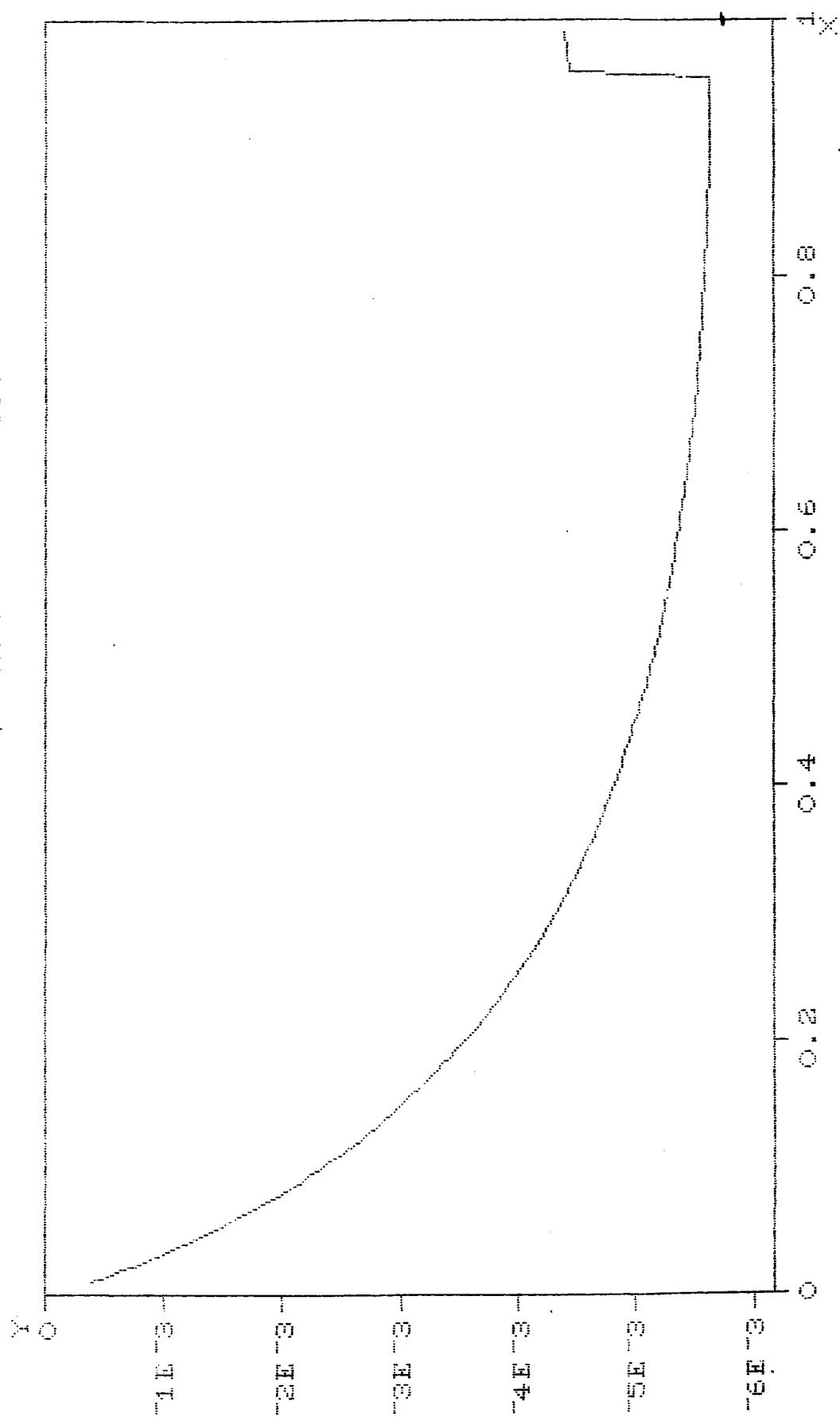
PLOT OF BSMIN AGAINST LAMBDA, PARAMETERS:1 1000



PLOT OF  $B_{MIN}$  AGAINST  $\lambda$ , PARAMETERS: 1 10



PLOT OF BSMIN AGAINST LAMEDA, PARAMETERS:1 0.1





#### 5.4.5 Mispecification of $\lambda$

Another interesting feature which is worth mentioning is the problem of using a model with  $\lambda = 1$  when the true value of  $\lambda$  lies in  $(0, 1)$ . We know that when we aggregate the observations, the coefficient in the system equation decays exponentially as  $\lambda^n$ , while the assumed model will maintain  $\lambda = 1$ . We want to know how this affects the one step ahead forecast error.

We adopt the same method that was used to derive (5.144). We still have (5.141) and (5.142) now is slightly modified as

$$e_k = Z_k - m_{k-1}$$

and we apply  $(1 - B)$  to the last equality to obtain

$$(1 - B)e_k = (1 - B)Z_k - (1 - B)m_{k-1} = (1 - B)Z_k - (1 - \beta)Be_k$$

where  $\beta$  is the discount factor to be used in the simple model (with  $\lambda = 1$ ). This gives

$$(1 - \beta B)e_k = (1 - B)Z_k$$

We combine this to (5.141) to obtain

$$e_k = (1 - \beta B)^{-1}(1 - B)(1 - hB)^{-1}(1 - \beta^* B)a_k \quad (5.155)$$

Then, again, if we write  $e_k = \phi(B)a_k$ , we will have  $\gamma_e(B) = \phi(B)\phi(B^{-1})\gamma_a(B)$ . This together with (5.155) gives

$$\gamma_e(B) = (1 - B)(1 - B^{-1})(1 - \beta^* B)(1 - \beta^* B^{-1})D(B)V(a) \quad (5.156)$$

where  $V(a)$  is the variance of  $a_k$  and  $D(B)$  is given by

$$\begin{aligned} D(B) &= \left( \sum_{i=0}^{\infty} \beta^i B^i \right) \left( \sum_{i=0}^{\infty} \beta^i B^{-i} \right) \left( \sum_{i=0}^{\infty} h^i B^i \right) \left( \sum_{i=0}^{\infty} h^i B^{-i} \right) \\ &= \frac{1}{(1 - \beta^2)(1 - h^2)} \left( \sum_{i=-\infty}^{\infty} \beta^{|i|} B^i \right) \left( \sum_{i=-\infty}^{\infty} h^{|i|} B^i \right) \\ &= \frac{1}{(1 - \beta^2)(1 - h^2)} \sum_{i=-\infty}^{\infty} c_{|i|} B^i \end{aligned} \quad (5.157)$$

In particular we have

$$c_0 = \frac{1 + h\beta}{1 - h\beta} \quad (5.158)$$

$$c_1 = \frac{h + \beta}{1 - h\beta} \quad (5.159)$$

$$c_2 = \frac{h^2 + \beta^2}{1 - h\beta} + h\beta \quad (5.160)$$

Now, rewrite (5.156) as

$$\frac{\gamma_e(B)}{V(a)} = (\beta^* B^2 - (1 + \beta^*)^2 B + 2(1 + \beta^* + (\beta^*)^2) - (1 + \beta^*)^2 B^{-1} + \beta^* B^{-2}) D(B)$$

Using (5.157) to (5.160), we can calculate the term in  $B^0$  in the above expression, obtaining

$$R(\beta) = \frac{V(e)}{V(a)} = \frac{2(1 + (\beta^*)^2) + \beta^*(1 - h) - (1 + h)\beta^*\beta}{(1 + h)(1 + \beta)(1 - h\beta)} \quad (5.161)$$

From (5.146) and (5.147),  $1 + (\beta^*)^2 = p\beta^*$ , where  $p = -\gamma_\eta(0)/\gamma_\eta(1)$ . Using this, define  $S(\beta)$  writing (5.161) as

$$R(\beta) = \frac{2\beta^*(p + 1 - h - (1 + h)\beta)}{(1 + h)(1 + \beta)(1 - h\beta)} = \frac{2\beta^*}{1 + h} S(\beta)$$

and work with  $S(\beta)$  instead of  $R(\beta)$ .

The derivative  $S'(\beta)$  has its numerator  $N(\beta)$  given by

$$N(\beta) = -h(1 + h)\beta^2 + 2h(p + 1 - h)\beta - (1 + h) - (1 - h)(p + 1 - h)$$

a concave quadratic function of  $\beta$ . Consider the case for which  $\beta^*$  is positive, which means  $p$  is positive. The product of the roots of the above polynomial is

$$\frac{1}{h} + \frac{(1 - h)(p + 1 - h)}{h(1 + h)} > \frac{1}{h} > 1$$

which means we must have no more than one root of  $N(\beta)$  in  $(0, 1)$ . But, observe that

$$N(0) = -1 - h - (1 - h)(p + 1 - h) < 0$$

Therefore, the existence of a root of  $N(\beta)$  in  $(0, 1)$  is simply determined by the sign of  $N(1)$ , which, in turn, is given by

$$N(1) = -(1 + h)^2 + (3h - 1)(p + 1 - h)$$

There must be a root in  $(0, 1)$  if and only if  $N(1)$  is positive, which is equivalent to

$$-\gamma_\eta(1) < \frac{3h - 1}{4h^2 - 2h + 2}$$

Because  $N(0) < 0$ , this root must correspond to a minimum point of  $R(\beta)$ .

When  $N(1) < 0$ , then  $S'(\beta)$  is always negative, which means our best choice is  $\beta = 1$ .

When  $\beta^* < 0$ , we can see that  $N(\beta)$  will be a decreasing function in  $(0, 1)$ . The behaviour of  $N(0)$  and  $N(1)$  will depend on each specific case. For illustration, let's consider an example with  $\lambda = 0.99$ . Suppose  $W = 10^4 V$ . For  $n = 3$ , we already have  $\beta^* < 0$ . In this case,  $N(0) < 0$ , and the best choice will, then, be  $\beta = 0$ . For reasonable values of  $n$  (around 60), we have  $N(0) > 0$  and  $N(1) < 0$ . Here, we compare  $R(0)$  and  $R(1)$  to see what is the smallest, and the answer is that we should choose  $\beta = 0$ . For a large value of  $n$ , (after 200), we have  $N(1) > 0$ , which means we must fix  $\beta = 1$ . Now, suppose  $W = 100V$ . In this case,  $\beta^*$  becomes negative for  $n = 25$ , and we begin with  $N(0) > 0$  and  $N(1) < 0$ . Comparing  $R(0)$  and  $R(1)$  we will conclude that  $\beta = 0$  must be chosen for relatively small values of  $n$  and  $\beta = 1$  must be chosen for relatively large values of  $n$ . For very large values of  $n$  (after 160), we will have  $N(1) > 0$ , and, therefore,  $\beta = 1$  must be chosen. Now, suppose  $W = V$ . In this case, as soon as  $\beta^*$  becomes negative, we get  $N(1) > 0$ , and the best decision will be to choose  $\beta = 1$ .

We can try to analyse the variation of the best choice for  $\beta$  as we increase the level of aggregation  $n$ . Suppose, to clarify things that  $\lambda \approx 1$ . For a reasonable value of  $n$ , we can still have  $\lambda^n \approx 1$  and the choice of  $\beta = 0$  can be justified by saying that the two clusters are so far apart that the best decision in order to forecast the next observation is to use only the last information. But, if we consider a very high aggregation level, then, in this case, we cannot consider  $\lambda^n \approx 1$  and we will also have a quick loss of information. Then, we come to a point where the information of the last observation does not mean anything to forecast the next, in other words, we must choose  $\beta = 1$ . We can also think how the relation between  $W$  and  $V$  influences the choice of  $\beta$ . If  $W \gg V$ , then the systematic effects almost completely dominate; we have a very high squared correlation between the state and the observation. Therefore, for a reasonable (but not very high) level of aggregation, we can still put the simple random walk as the best choice, in the sense that  $Z_{k-1}$  should be everything we use to forecast  $Z_k$ . On the other hand, if  $W$  is comparable to  $V$ , then, the accumulated observational errors will dominate when we aggregate, and this will bring a large loss of information. Therefore, we must very quickly get to the point where the best decision is to use the model where  $Z_{k-1}$  does not bring any information to forecast  $Z_k$ , which means to put  $\beta = 1$ .

## 5.5 General problem and linear growth

### 5.5.1 General idea

Now we try to extend the ideas we explored for the very simple case of a univariate system equation to a more general model, where we have a state vector. The general

idea can be summarized as follows. Suppose  $y_t$  evolves according to (5.56) and (5.57) and let  $p$  be the dimension of the state vector. Assume without loss of generality that  $G$  is in Jordan form. Let  $\lambda_1, \dots, \lambda_p$  be the eigenvalues of  $G$  and define  $\phi(B) = (1 - \lambda_1 B) \cdots (1 - \lambda_p B)$ , where  $B$  is the backshift operator. Then, it is readily seen from (5.57) that  $\phi(B)\theta_t$  is a pure  $MA(p-1)$  vector process, and, from (5.56), we can conclude that  $y_t$  is an *ARIMA* process. Now, consider the aggregated process  $Z_k$ , defined as in (5.1). We first observe that, because we are aggregating each  $n$  units of time in one observation, application to the left side of (5.1) of the  $B$  operator (that means, backshift with respect to  $k$ ) is the same as applying  $B^n$  to the right side (lag- $n$  backshift with respect to  $t$ ). Let's consider, then, the operator  $\Phi(B) = (1 - \lambda_1^n B) \cdots (1 - \lambda_p^n B)$ . We have, for example,

$$\Phi(B)Z_{p+1} = \sum_{i=1}^n (1 - \lambda_1^n B^n) \cdots (1 - \lambda_p^n B^n) y_{np+i}$$

Now, observe that each  $(1 - \lambda_j^n B^n)$  above can be factored like  $(1 + \lambda_j B + \cdots + \lambda_j^{n-1} B^{n-1})(1 - \lambda_j B)$ . Therefore, we can write

$$\Phi(B)Z_{p+1} = \sum_{i=1}^n \psi(B)\phi(B)y_{np+i} \quad (5.162)$$

where

$$\psi(B) = \prod_{j=1}^p (1 + \lambda_j B + \cdots + \lambda_j^{n-1} B^{n-1})$$

If we write (5.162) as a linear combination of the  $v$ 's and  $\omega$ 's, we can see that lags of  $v$ 's will run from 1 to  $n(p+1)$  and lags of  $\omega$ 's will run from  $np+1-(p-1)-(n-1)p = 2$  to  $n(p+1)$ . Hence, the process  $\eta_k = \Phi(B)Z_{p+k}$  will be such that  $E[\eta_k \eta_{k+j}] = 0$  for  $j > p$ . In other words,  $Z_k$  has an *ARIMA* structure, where the roots of the autoregressive polynomial are  $\lambda_1^{-n}, \dots, \lambda_p^{-n}$ . We write

$$Z_k = \sum_{j=1}^p \psi_j Z_{k-j} + \sum_{j=1}^p \pi_j a_{k-j} + a_k$$

where  $a_k$  represents an uncorrelated process. Then, defining the  $p+1$  dimensional vectors  $F^* = (1, 0, \dots, 0)$  and  $\delta_k = (1, \pi_1, \dots, \pi_p)a_k$ , together with the  $(p+1) \times (p+1)$  matrix  $G^*$ , given by

$$G^* = \begin{pmatrix} \psi_1 & 1 & 0 & \dots & 0 \\ \psi_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_p & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

provides the well-known state space representation for  $Z_k$  (see West and Harrison, 1989), defined by the equations

$$\begin{cases} Z_k = F^{*'} \phi_k \\ \phi_k = G^* \phi_{k-1} + \delta_k, \end{cases}$$

It can be readily seen that eigenvalues of  $G^*$  are  $\lambda_1^n, \dots, \lambda_p^n$  and 0. We immediately conclude, therefore, that the inclusion of a zero eigenvalue in the system matrix will always give a solution for the representation problem, because we have, at least, the well-known state space representation above defined (we suppose the initial information is good enough, so we don't have to worry about it). An interesting question concerns the conditions for which we can simplify by considering a  $p$ -dimensional DLM representation, where the system matrix will have eigenvalues  $\lambda_1^n, \dots, \lambda_p^n$ , without having to include a zero eigenvalue.

### 5.5.2 The linear growth model

We turn now to the linear growth model, trying to extend what has been done before to a more sophisticated example. We consider, then, a process  $y_t$ , evolving in time according to (5.56) and (5.57), where  $F = (1, 0)'$  and  $G = J_2(1)$ , a two-dimensional Jordan block with 1's in its diagonal. We rewrite the system equation in the more convenient way below:

$$\mu_t = \mu_{t-1} + \beta_t + \delta\mu_t \quad (5.163)$$

$$\beta_t = \beta_{t-1} + \delta\beta_t \quad (5.164)$$

where  $Var[\delta\mu_t] = W_\mu$ ,  $Var[\delta\beta_t] = W_\beta$  and  $Cov[\delta\mu_t, \delta\beta_t] = W_{\mu\beta}$ .

Then, if data is aggregated each  $n$  observations to obtain  $Z_k$ , we try to represent  $Z_k$  by the similar form:

$$Z_k = \mu_k^* + \epsilon_k \quad (5.165)$$

$$\mu_k^* = \mu_{k-1}^* + \beta_{k-1}^* + \omega_{1,k}^* \quad (5.166)$$

$$\beta_k^* = \beta_{k-1}^* + \omega_{2,k}^* \quad (5.167)$$

where all errors have zero mean,  $Var[\epsilon_k] = V^*$ ,  $Var[\omega_{1,k}^*] = W_1^*$ ,  $Var[\omega_{2,k}^*] = W_2^*$ ,  $Cov[\omega_{1,k}^*, \omega_{2,k}^*] = W_{12}^*$ , and  $\epsilon_k$  is uncorrelated with the errors in the system equation. We will use essentially the same ideas we have applied before, considering the process  $\eta_k = \Delta^2 Z_k$ , where  $\Delta = 1 - B$ . Using (5.165) to (5.167), we get

$$\eta_k = \omega_{2,k-1}^* + \Delta\omega_{1,k}^* + \Delta^2\epsilon_k$$

Therefore, denoting again by  $\gamma_\eta(j)$  the autocovariance of  $\eta$  of lag  $j$ , we obtain

$$\gamma_\eta(0) = 6V^* - 2W_{12}^* + 2W_1^* + W_2^* \quad (5.168)$$

$$\gamma_\eta(1) = -4V^* + W_{12}^* - W_1^* \quad (5.169)$$

$$\gamma_\eta(2) = V^* \quad (5.170)$$

Now, we calculate the autocovariances above from the original model, and by comparison, we try to obtain valid conditions to have representation (5.165)-(5.167). For this purpose, let's consider the third aggregated observation,  $Z_3 = y_{1+2n} + \dots + y_{3n}$ . We have

$$\begin{aligned} \Delta^2 Z_3 &= \sum_{i=1}^n (y_{2n+i} - 2y_{n+i} + y_i) \\ &= \sum_{i=1}^n (\mu_{2n+i} - 2\mu_{n+i} + \mu_i) + \sum_{i=1}^n (v_{2n+i} - 2v_{n+i} + v_i) \end{aligned}$$

Now, use (5.163) and (5.164) to get

$$\begin{aligned} \mu_{2n+i} - \mu_{n+i} &= (\beta_{n+i+1} + \dots + \beta_{2n+i}) + (\delta\mu_{n+i+1} + \dots + \delta\mu_{2n+i}) \\ \mu_{n+i} - \mu_i &= (\beta_{i+1} + \dots + \beta_{n+i}) + (\delta\mu_{i+1} + \dots + \delta\mu_{n+i}) \end{aligned}$$

Therefore

$$\begin{aligned} \mu_{2n+i} - 2\mu_{n+i} + \mu_i &= \sum_{j=1}^n (\beta_{n+i+j} - \beta_{i+j}) + \sum_{j=1}^n \delta\mu_{n+i+j} - \sum_{j=1}^n \delta\mu_{i+j} \\ &= \sum_{j=1}^n \sum_{k=1}^n \delta\beta_{i+j+k} + \sum_{j=1}^n \delta\mu_{n+i+j} - \sum_{j=1}^n \delta\mu_{i+j} \end{aligned}$$

Hence

$$\begin{aligned} \Delta^2 Z_3 &= \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \delta\beta_{i+j+k} + \sum_{i=1}^n \sum_{j=1}^n \delta\mu_{n+i+j} - \sum_{i=1}^n \sum_{j=1}^n \delta\mu_{i+j} \\ &\quad + \sum_{i=1}^n v_{2n+i} - 2 \sum_{i=1}^n v_{n+i} + \sum_{i=1}^n v_i \end{aligned}$$

In order to calculate the autocovariances of  $\eta_k$  using the above expression, we need to know, for each error, how many times each lag appears in the sum. The coefficients related to  $\delta\beta$  are a bit harder to obtain, the coefficient for time  $p$  corresponding to the number of positive integer solutions of the equation  $i + j + k = p$ , with the

restriction  $i, j, k \leq n$ . The answer turns out to be

$$\begin{aligned}\Delta^2 Z_3 = & \sum_{p=3}^{n+2} A_p \delta \beta_p + \sum_{p=n+3}^{2n} B_p \delta \beta_p + \sum_{p=2n+1}^{3n} C_p \delta \beta_p + \\ & \sum_{p=2}^{n+1} D_p \delta \mu_p + \sum_{p=n+2}^{2n} E_p \delta \mu_p + \sum_{p=2n+1}^{3n} F_p \delta \mu_p + \\ & \sum_{p=1}^n v_p - 2 \sum_{p=n+1}^{2n} v_p + \sum_{p=2n+1}^{3n} v_p\end{aligned}$$

where

$$A_p = \frac{1}{2}(p-1)(p-2) \quad (5.171)$$

$$B_p = -p^2 + 3(1+n)p - \frac{1}{2}(3n^2 + 9n + 4) \quad (5.172)$$

$$C_p = \frac{1}{2}(3n+2-p)(3n+1-p) \quad (5.173)$$

$$D_p = 1 - p \quad (5.174)$$

$$E_p = 2p - 2 - 3n \quad (5.175)$$

$$F_p = 3n + 1 - p \quad (5.176)$$

From this, we can calculate the first three autocorrelations of  $\eta_k = \Delta^2 Z_k$ . We get

$$\begin{aligned}\gamma_\eta(0) = & \left\{ \sum_{p=3}^{n+2} A_p^2 + \sum_{p=n+3}^{2n} B_p^2 + \sum_{p=2n+1}^{3n} C_p^2 \right\} W_\beta + \\ & \left\{ \sum_{p=2}^{n+1} D_p^2 + \sum_{p=n+2}^{2n} E_p^2 + \sum_{p=2n+1}^{3n} F_p^2 \right\} W_\mu + \\ & 2 \left\{ \sum_{p=3}^{n+1} A_p D_p + A_{n+2} E_{n+2} + \sum_{p=n+3}^{2n} B_p E_p + \sum_{p=2n+1}^{3n} C_p F_p \right\} W_{\mu\beta} + 6nV \quad (5.177)\end{aligned}$$

$$\begin{aligned}\gamma_\eta(1) = & \left\{ \sum_{p=3}^n A_p B_{p+n} + \sum_{p=n+1}^{n+2} A_p C_{p+n} + \sum_{p=n+3}^{2n} B_p C_{p+n} \right\} W_\beta + \\ & \left\{ \sum_{p=2}^n D_p E_{p+n} + D_{n+1} F_{2n+1} + \sum_{p=n+2}^{2n} E_p F_{p+n} \right\} W_\mu + \\ & \left\{ \sum_{p=3}^n A_p E_{p+n} + \sum_{p=n+1}^{n+2} A_p F_{p+n} + \sum_{p=n+3}^{2n} B_p F_{p+n} \right\} W_{\mu\beta} + \\ & \left\{ D_2 A_{n+2} + \sum_{p=3}^n D_p B_{p+n} + D_{n+1} C_{2n+1} + \sum_{p=n+2}^{2n} E_p C_{p+n} \right\} W_{\mu\beta} - 4nV \quad (5.178)\end{aligned}$$

$$\gamma_\eta(2) = \left\{ \sum_{p=3}^n A_p C_{p+2n} \right\} W_\beta + \left\{ \sum_{p=2}^n D_p F_{p+2n} \right\} W_\mu + \left\{ \sum_{p=3}^n A_p F_{p+2n} + \sum_{p=2}^n D_p C_{p+2n} \right\} W_{\mu\beta} + nV \quad (5.179)$$

Then, using (5.171) to (5.176) we can calculate the sums in (5.177) to (5.179) in order to obtain the final expressions for the first three autocorrelations of  $\eta_k$ . After a calculation, we finally arrive at

$$\begin{aligned} \gamma_\eta(0) &= \frac{n(11n^4 + 5n^2 + 4)}{20} W_\beta + n(n^2 + 1) W_\mu + n(2n^2 - n + 1) W_{\mu\beta} + 6nV \\ \gamma_\eta(1) &= \frac{n(n^2 - 1)(13n^2 + 8)}{60} W_\beta - \frac{n(n^2 + 2)}{3} (W_\mu + W_{\mu\beta}) - 4nV \\ \gamma_\eta(2) &= \frac{n(n^2 - 1)(n^2 - 4)}{30} W_\beta - \frac{n(n^2 - 1)}{6} (W_\mu + W_{\mu\beta}) + nV \end{aligned}$$

We compare these last three expressions with (5.168) to (5.170) in order to obtain the necessary conditions for the existence of a DLM representation of the aggregated data. It is interesting to observe that from (5.168) to (5.170) we can get expressions for  $V^*$ ,  $W_{12}^* - W_1^*$ , and  $W_2^*$ . This seems to indicate that the variance in the first system equation can be probably chosen within a certain degree of freedom in our representation (if we choose the covariance between the errors accordingly) while the variance in the second equation is completely determined by the autocovariance structure of the aggregated series. From the comparison above described, we will get

$$V^* = nV - \frac{n(n^2 - 1)}{6} (W_\mu + W_{\mu\beta}) + \frac{n(n^2 - 1)(n^2 - 4)}{30} W_\beta \quad (5.180)$$

$$W_1^* - W_{12}^* = n^3 (W_\mu + W_{\mu\beta}) - \frac{n(n^2 - 1)(7n^2 - 8)}{20} W_\beta \quad (5.181)$$

$$W_2^* = n^2(n - 1) W_{\mu\beta} + \frac{n(21n^4 - 5n^2 + 4)}{20} W_\beta \quad (5.182)$$

The quantites above must, in our representation, satisfy  $V^* \geq 0$ ,  $W_2^* \geq 0$  and  $(W_{12}^*)^2 \leq W_1^* W_2^*$ , the two last conditions making sure that the  $W^*$  matrix is semi-positive definite. Let  $M^* = W_1^* - W_{12}^*$ . Then, our last condition can be rewritten as

$$(W_{12}^*)^2 \leq W_2^* (M^* + W_{12}^*)$$

Therefore

$$(W_{12}^*)^2 - W_2^* W_{12}^* - M^* W_2^* \leq 0 \quad (5.183)$$

and this can only be possible if the discriminant  $\Delta = b^2 - 4ac$  in the above inequality is non-negative. Since

$$\Delta = W_2^* (W_2^* + 4M^*)$$



our conditions can be rewritten as  $W_2^* \geq 0$  and  $W_2^* + 4M^* \geq 0$ . This last quantity can be written as

$$W_2^* + 4M^* = 4n^3W_\mu + n^2(5n-1)W_{\mu\beta} - \frac{n(7n^4 - 55n^2 + 28)}{20}W_\beta \quad (5.184)$$

Then, (5.180), (5.182) and (5.184) must be non-negative to guarantee the representation we are looking for. Observe, as well, that, when these conditions are satisfied, we can choose  $W_{12}^* = 0$  if and only if the product of the roots in the left side of (5.183) is negative, which means,  $M^* > 0$ .

Consider the standard DLM representation, given by

$$\begin{aligned} \mu_t &= \mu_{t-1} + \beta_{t-1} + \omega_{1,t} \\ \beta_t &= \beta_{t-1} + \omega_{2,t} \end{aligned}$$

where  $Var[\omega_{1,t}] = W_1$ ,  $Var[\omega_{2,t}] = W_2$  and  $Cov[\omega_{1,t}, \omega_{2,t}] = W_{12}$ . Comparing this with (5.163) and (5.164), we get  $W_1 = W_\mu + W_\beta + 2W_{\mu\beta}$ ,  $W_2 = W_\beta$  and  $W_{12} = W_\beta + W_{\mu\beta}$ . We can, then, rewrite (5.180) to (5.182) in terms of these quantites, as

$$V^* = nV - \frac{n(n^2 - 1)}{6}(W_1 - W_{12}) + \frac{n(n^2 - 1)(n^2 - 4)}{30}W_2 \quad (5.185)$$

$$W_1^* - W_{12}^* = n^3(W_1 - W_{12}) - \frac{n(n^2 - 1)(7n^2 - 8)}{20}W_2 \quad (5.186)$$

$$W_2^* = n^2(n-1)W_{12} + \frac{n(21n^4 - 25n^2 + 20n + 4)}{20}W_2 \quad (5.187)$$

It may be important to consider some particular cases of the results obtained above. The first thing we can see is that (5.185) to (5.187) are consistent with the anterior case, where the system equation were unidimensional. In this particular subcase we have  $W_2 = W_{12} = 0$ . Then, (5.187) gives us  $W_2^* = 0$  which implies  $W_{12}^* = 0$ . Using this in (5.186) gives us  $W_1^* = n^3W_1$ . Also, (5.185) will be coincident with (5.7). Observe that the same conclusion can be drawn if we have  $W_\beta = W_{\mu\beta} = 0$  in (5.180) to (5.182).

Another interesting subcase corresponds to uncorrelated errors in the system equation. If  $W_{12} = 0$ , we can readily see, from (5.187), that  $W_2^* > 0$ , hence, we do not have to worry about this condition. Now, let's consider the condition on determinant of  $W^*$ , given by  $W_2^* + 4M^* > 0$ . We have

$$W_2^* + 4M^* = 4n^3W_1 - \frac{n(7n^4 - 35n^2 - 20n + 28)}{20}W_2 - n^2(3n+1)W_{12}$$

and the condition reduces to

$$4n^3W_1 - \frac{n(7n^4 - 35n^2 - 20n + 28)}{20}W_2 \geq 0$$

It is interesting to see that for  $n \geq 3$  the veracity of this condition will depend on the relation between  $W_1$  and  $W_2$  (for  $n = 3$ , for example, it will be written as  $108W_1 - 33W_2 \geq 0$ ), but, for  $n = 2$  it is reduced to  $32W_1 + 4W_2 \geq 0$ , which is always satisfied. This means to say we just have to worry about  $V^* \geq 0$ , when  $n = 2$  in the uncorrelated case. But, looking at (5.185), we observe that the coefficient of  $W_2$  vanishes, due to the factor  $n^2 - 4$ , when  $n = 2$ . Therefore, the condition is just reduced to that one we had in the first order polynomial DLM, namely

$$V \geq \frac{W}{2}$$

In other words, the uncorrelated system has no sensitivity to the linear growth structure when we aggregate for only two observations. The complexity of the linear growth structure will only lead to the other two extra conditions from the moment we aggregate more than two observations each time.

An interesting feature which deserves to be mentioned is the fact that the aggregated data can sometimes be represented by a model which is simpler than the original one. A trivial example can be given when  $M^* > 0$ . In this case, we can represent the aggregated series by a linear growth structure where the errors in the system equation are uncorrelated. That means, we can have a linear growth structure where the errors are correlated and end up with aggregated data which can be represented by a more simple linear growth model, with uncorrelated errors in the system equation. Another interesting example is the one that follows. Suppose  $V = 30$ ,  $W_1 = 53$ ,  $W_{12} = -7$  and  $W_2 = 1$ . We aggregate the series for  $n = 2$  observations. Then, from (5.185) and (5.187) we will get  $W_2^* = 0$  and  $V^* = 0$ . Therefore, although the original data was represented by that complex linear growth structure, the aggregated series  $Z_k$  can be represented by the more simple model

$$Z_k = \beta + Z_{k-1} + \omega_k$$

where  $\beta$  is a constant increment and  $\omega_k$  is a random noise. Observe that from (5.186) we also get  $Var[\omega_k] = 474$ , which is a high variance, compared with the others we give in the original model. This can be explained from the fact that all the variability of the aggregated series is thrown into the single random component of the reduced model we have for the aggregated data.

## 5.6 Aggregation and the booking model

We consider again the model discussed in the last chapter. As we have stated before, a very important aspect of aggregation lies in the fact that some factors that have importance in explaining a given quantity may lose importance in the aggregated level, and, at this level, other effects may become much more important. In time series, aggregation has the effect of smoothing the high frequencies out, acting as a high frequency filter. At a very low level of aggregation, that is, if data is sampled too frequently, a model fitted to explain data variation will capture mainly the high frequency effects. The low frequency components will not play an important part of the model, and will probably not even be well identified when we try to explain the total variation of data. Therefore, it is very important to state *a priori* what our main interests are when we are to build a model, and the level of aggregation plays a fundamental role in the precision we want. Also, if the qualitative form of the model we have fixed for the original data is correct, then, we expect a decrease in the discount factor when we aggregate. If the discount factor is increasing with aggregation, then the qualitative form we are using is more appropriate at higher levels of aggregation. At these higher levels of aggregation, the higher frequencies that exist at low levels are increasingly diminishing and the systematic variation becomes dominant.

In the case of the booking model, one week can be a very small interval of time if we want to capture the long term behaviour of the booking process. The number of booked tickets can be influenced by variables such as promotions, reduction in the price or even weather. Hence, if we are interested in the long term behaviour of our data, it may be worth considering aggregated data, instead. It is important to observe that the booking problem presents two dimensions of time. The first dimension is that one related to the booking interval. It is easy to extend the model to consider aggregated data in this dimension, since the sum of independent Poisson distributions is also Poisson distributed. The second dimension is that one referent to the time interval between consecutive flights. Study of aggregation along this time dimension is much more difficult. In this case we have to consider the distribution of the sum of correlated negative binomial distributions, the correlation being calculated through the defined bijection between the gamma and log-normal distributions.

The extension of the booking model in order to allow for aggregation of different weeks in the same flight is straightforward. Remember we had an evolving parameter  $\eta_t$ , such that, conditional on  $\eta_t$ , the variables  $x_{1t}, \dots, x_{kt}$  were indepen-

dently and Poisson distributed with mean  $r_i \exp(\eta_i)$ , where  $r_i$  defines the reservation curve. Consequently, the aggregated data will still be independently and Poisson distributed and we aggregate the reservation curve accordingly. In order to allow for aggregation, we modify the updating procedure in such a way that the updating is made each  $n$  periods of time, where  $n$  corresponds to the aggregation level. The reason is that if the updating is modified like this, then, when updating the model we will have information about all the flights. Then, the two dimensions of time are kept equal, i.e., the time interval for which we observe is equal to the time interval for which we update. A third feature is that the updating vector for the  $\lambda$ 's will be reduced to  $k - n + 1$  dimensions, where, again,  $k$  corresponds to the number of weeks in the booking period and  $n$  is the aggregation level. This is because when we work with the aggregated data, we have no observations for the last  $n - 1$  flights. It is important to observe that aggregation in this dimension also contributes to smooth the high frequencies out. This can be explained from the fact that if we have Poisson distributions with very small parameters, then, data along time will be subject to much variation, which can be eliminated as soon as we aggregate two or more weeks.

## 5.7 Conclusions

The fundamental idea in this chapter is the importance of the sampling interval in forecasting. When data is sampled too frequently, the high frequency components will dominate. These components are not difficult to characterise and bring unnecessary complications if we want to forecast the long term behaviour of the process. Data aggregation can reduce and smooth the high frequencies out, making it easier for the modeller to construct an effective model, which is of great importance if we are interested in long term forecasting.

When we put this study in the DLM framework, it becomes important to know how the aggregated data can be represented as a DLM process, if the original data admits this representation. We recall that the presence of high frequencies in such data can be characterized by the relation between the observational and systematic variances. If the observational variance is much bigger than the systematic ones, then, we expect the high frequency components to become dominant. If, otherwise, the observational variances are much lower, then, we expect the lower frequencies to dictate the process behaviour.

Our first result (Theorem 1) is concerned with the very simple univariate random walk plus noise. We have shown that if the two conditions given by (5.4) are satisfied, then, the aggregated series can still be represented by a simple random walk plus noise model. The first condition is the structural one, the second being

related to the initial information about the state parameter. From (5.7) and (5.8), it is interesting to observe that, for the aggregated model, the relation between the variances becomes

$$\frac{V^*}{W^*} = \frac{1}{n^2} \left( \frac{V}{W} - \frac{n^2 - 1}{6} \right)$$

and it is clear from above that this decreases with  $n$ . We then expect the discount factor to decrease, since the systematic variance becomes dominant with aggregation, and the weight given to the observation must therefore be higher in the updating equations.

When conditions are not satisfied the solution is to introduce a correlation between the state parameter and the observational error in the observation equation. Equivalently, we introduce a zero eigenvalue in the system matrix. Then, from Theorem 2, we can have a set of possible covariances between the state and error, which allows us to represent the aggregated data by an extended model. For a very high aggregation level we need to introduce a high correlation between the errors to have a valid DLM representation. But, in terms of forecasting, we do not lose very much in not considering this correlation (see 5.151), in the sense that the one step ahead error variance will be increased by a relatively low factor. We also concluded that the best discount factor to be used must be zero, after restriction is broken, and this means that after this threshold things are smoothed out to a point where the best simple model to be used (i.e., not considering the correlation) is the simple random walk model, and this has a performance almost as good as the sophisticated model (with the correlation introduced).

The next simplest model is that one given by (5.58) and (5.59). For this model, we also derived a set of conditions that, if satisfied, then the aggregated series can be represented by the similar model (5.60) and (5.61) and these conditions reduce to those for the anterior model, in the particular case when  $\lambda = 1$ . If the conditions are not satisfied, we again introduce a zero eigenvalue, and we can have a DLM representation for the aggregated data. Again, for a high aggregation level, the correlation between the residuals must be quite large, in order to allow for the representation. But, for  $\lambda > 0$ , it seems again that, in terms of forecasting, we do not lose much by not considering such correlation, since we can be almost as efficient as in the sophisticated model. Again, the best choice is to consider the simple random walk representation, in this case.

In the general case, where the state space vector is  $k$ -dimensional, the zero eigenvalue solution will always allow us to have a representation for the aggregated data. The major problem is still to know when this representation can be made simple,

i.e., when we can maintain the same dimension for the state space vector, without needing to increase it by one. If  $k$  is the dimension of the state space vector, then, we must satisfy  $k + 1$  conditions of positivity, corresponding to the  $k$  principal minors of the covariance matrix in the system equation, which must be positive definite, plus the positivity of the observational variance. We have studied the linear growth case with unit eigenvalues. It is important to observe that aggregation can work to simplify the structural form of the model, as we have seen in the numerical example we gave. This shows, then, that as we probably need a more sophisticated model when data is aggregated, we can also obtain a simplification of the model with aggregation. This kind of behaviour make us believe that most part of time there must be an optimal level, i.e., there must be an optimal sampling interval that we must use, in order to obtain the simplest possible model consistent with the forecasting objectives.

When applied to the booking problem, aggregation seems to be a good idea if we want to detect the long term behaviour for the booking of some particular flights. The period of one week time seems to be quite small for some flights, in the sense that there must be a lot of variation in weekly data, due to a number of factors, like advertising campaigns, reduction of prices, etc. Therefore, for these flights, it is recommended that we deal with the aggregated model, instead, such that we can eliminate these high frequency variations. The dimension of the state space vector can probably be diminished in practice, since we will probably need less variables for explanation of the trend when we aggregate these flights for more than one week. Two types of aggregation can be used here. The first one is the aggregation along the booking period of same flights, which is easily dealt with in our model, since we work with independent Poisson distributions. This can contribute to filter high variations, when we aggregate very small numbers. The second one is the aggregation of data concerning with different flights. This is a more difficult approach, but a very important one, since it contributes to eliminate most of the cross flight effects, and it is left for future investigation.

## CHAPTER 6

### CONCLUSIONS

#### 6.1 Review of Chapter 3

In Chapter 3 we studied the reference analysis of the DLM. The proposed approach is an improvement on that of Pole and West, which is presented in Section 2.6. It uses the initial observations much more efficiently, in the sense that it uses the information postulated by the model in order to derive an initial proper distribution for a subspace of the state space. This distribution is derived by considering the possibility that some of the eigenvalues of the system matrix may lie in the interior of the unit circle of the complex plane. That is, a proper marginal prior for stationary time series components can always be postulated based only upon the model form.

Also, we take into account the fact that, even when we cannot find a proper distribution for the entire state space, this is of no importance, as long as we have a proper prior for that subspace in which we wish to forecast. If this is the case, then, we can obtain a proper *conditional* forecast distribution for the future observations of the series. Giving an example, if a variable such as temperature has never varied from, say,  $100^{\circ}\text{C}$ , we can make forecasts as long as temperature remains at this specific value. But, if conditions change to, say,  $110^{\circ}\text{C}$ , a forecast is not possible, although after observing what happens, forecasts can be made for any future temperature change.

Another point that deserves to be mentioned is that our approach admits singularity of the system matrix, a situation that can occur in practice. This generalises the original work of Pole and West, which assumes the system matrix to be non-singular.

#### 6.2 Review of Chapter 4

In Chapter 4 we have developed a model for airline passengers booking tickets for weekly flights. An important reason for this study lies in the many drawbacks presented by the models that are classically used for this purpose. We consider the booking period (that interval of time between the beginning of bookings and departure of the flight) of  $k$  weeks and collect a set  $\{X_{it}, i = 1, \dots, k; t = 1, \dots, T\}$  of observations, where  $X_{kt}$  denotes the number of passengers booking seats in the plane during the  $k$ -th booking week for the flight departing at week  $t$ . Our model supposes that  $\{X_{1t}, \dots, X_{kt}\}$  are independent Poisson observations and  $X_{it} \sim Po(r_i \varphi_t)$ . The parameter  $\varphi_t$  corresponds to a mean demand level for the total number of booked tickets in the plane and the  $r_i$ 's define the reservation curve, subject to  $0 < r_i < 1, \forall i$  and  $\sum_{i=1}^n r_i = 1$ . The natural parameter  $\eta_t = \log(\varphi_t)$  of the Poisson distribution

is linked to another parameter  $\lambda_t$  through a known bijection  $\lambda_t = f(\eta_t)$ . This new parameter  $\eta_t$  is obtained from the  $n$ -dimensional state vector  $\theta_t$  through a linear function  $\lambda_t = F_t' \theta_t$ , the vector  $F_t$  being known for each instant of time. The state vector  $\theta_t$  itself is supposed normally distributed, evolving in time according to the usual equation  $\theta_t = \theta_{t-1} + \omega_t$ , where  $\{\omega_t\}$  is a sequence of independent errors. This puts our model in the Dynamic Generalised Linear Model framework. This model presents advantages over classical models used for the same purpose. It is a dynamical model, which permits the analyst to have control over the structural changes that can occur in data behaviour, as opposed to the classical models, which are static. It is also more realistic, since it appropriately assumes data to be Poisson distributed, given a certain parameter. Classical models presuppose normality, a hypothesis that is extremely inadequate for our purposes, since we are dealing with small integer numbers and expect to be working with skewed distributions.

In order to have conjugacy we assume that  $\varphi_t$  has a gamma distribution for each instant  $t$ . On the other hand, we assume that  $\theta_t$  is normally distributed, and, in consequence,  $\lambda_t$  is also. Therefore, we construct a bijection between the log-normal and gamma distributions, to play the role of the bijection  $f(\cdot)$  we have just cited in the above paragraph. This bijection is obtained from a minimization problem. Given a fixed gamma density  $p$ , we obtain a log-normal density  $q$  that approximately minimizes the  $L_2$  distance given by (4.13). We solve this problem numerically for some gamma distributions. In each problem, we fix a gamma distribution with unit mean and a given coefficient of variation  $k_G$ , and, then, find a log-normal that approximately minimizes the  $L_2$  distance between the curves. After doing this for some coefficients of variation, we obtained the relations (4.14.2) and (4.15), by considering a regression of the solutions on  $k_G$ .

After that, we develop the updating described in Sections 4.7 and 4.8, using also the concept of equivalent observation. For a simple test of the model, we have generated a random matrix  $X$ , where  $x_{ij}$  has a Poisson distribution with mean proportional to a parameter  $\varphi_j$ , this parameter being log-normally distributed. We, then, fit our model, which is slightly different, in the sense that we use a gamma distribution, instead of a log-normal, in order to have conjugacy. The adequacy of the model is tested with the goodness-of-fit statistics defined by (4.23). This is calculated using the fact that the model fits a negative binomial distribution for each observation and is tested against a  $N(0, 1)$  distribution (see details in Section 4.9).



### 6.3 Analysis of results

We first consider the gamma to log-normal approximation. It is seen that the method of equating mean and variance is very good when the coefficient of variation  $k$  is very small and good for ( $k \leq 0.25$ ). This result is expected, since, for very small values of the coefficient of variation, both densities can be very well approximated by a normal density. For larger values of the coefficient of variation of the gamma distribution, the relations (4.14.2) and (4.15) which defines our bijection are better than simply equating the first two moments. It is interesting to observe that the density approximation we have defined is such that the two curves are very near each other in the uniform sense. Therefore, our relationship performs a job which is almost as good as the sup norm minimization, which is a much more strong and difficult procedure. We also observe that continuity is a very important property of the defined relations, since we are interested in using a robust approach, in the sense that small perturbations in the gamma distribution must produce small perturbations in the parameters of the associated log-normal distribution.

Having produced this approximation, we apply it to the specific model we have defined. We first observe that, from standard conjugate analysis, the updated posterior mean  $\mu^*$  of the gamma distribution we fix for the Poisson parameter is given by

$$\mu^* = p\mu + qx \quad (6.1)$$

where  $\mu$  is the prior mean,  $x$  is the observation,  $p = \beta/(\beta + r)$ ,  $p + q = 1$  and we suppose the prior distribution is  $G(\alpha, \beta)$ . Suppose, now, that in the model evolution equation we use a discount factor which is very high (near one). That means we do not have a considerable change in the normal distribution of the state parameter, and, consequently, no considerable change in the log-normal distribution for  $\exp \lambda_t$ . From continuity of the bijection (and here again becomes clear the importance of continuity), we do not have a reasonable modification in the parameters of the gamma distribution of  $\varphi_t$ . In particular, we expect only a slight modification in the  $\beta$  parameter for this gamma distribution. Then, because of the form  $p$  is obtained in (6.1), we conclude that after a certain point we will have  $p$  almost equal to one, which means that virtually all weight will be given to the prior mean, and no much importance will be given to the observation. Now, suppose, on the other hand, that the discount factor is relatively small. Then, we expect a reasonable increase in the variance of the normal distribution for the state parameter. That means we expect a reasonable increase in the mean and coefficient of variation of the associated

log-normal. But, recalling (4.26), we have

$$\beta_G = \frac{0.2886 + k_G^{-2}}{\mu_L}$$

which shows that a high increase in  $\mu_L$  and  $k_L$ , implies, from (4.14.2), a high increase in  $\mu_L$  and  $k_G$ , and, consequently, a high decrease in  $\beta_G$ . If  $\beta_G$  becomes small, then, we will use (6.1) with a small  $p$  and the information from the observation will dominate. The conclusion is that we have preserved here the basic philosophy introduced in Chapter 2, that a large discount factor is translated as a simple model with an almost constant forecast function, the approximation becoming more local as the discount factor decreases. Observe that this conclusion is only possible because we have adopted a quadratic function of  $k_G$  in (4.15), and this guarantees that  $\beta_G$  decreases as  $k_G$  increases. If we have used a higher degree polynomial, a cubic polynomial, for example, we could not guarantee, in principle, that  $\beta_G$  would be a decreasing function of  $k_G$ .

We have tested our model with some simple simulations, which are briefly described in the last section. The forecasting performance for these cases can be summarized in the graphs (G 4.1) to (G 4.3). It can be seen that the mode of the negative binomial forecast distribution that is fitted to data is always very near the mean of the distribution. That means that the spot decision represented by the mode will be always coincident with the least squares decision represented by the mean, and our forecast will always be given by the most probable value. We recall that, in the negative binomial distribution, the mean  $\mu$  is linked to the mode  $n^*$  via the relation

$$n^* = \max_{\#} \left\{ 0, \left[ \left[ \mu - \frac{q}{p} \right] \right] \right\}$$

where  $\llbracket x \rrbracket$  denotes the integer part of  $x$ , and  $p, q$  are as in (6.1). Then, this result is expected, if we use a model with a large discount factor, and, consequently, with a small  $q/p$  ratio.

We have simulated a set of data, which are Poisson observations conditional on a parameter with a log-normal distribution. For this data, we have fitted our model, where we consider the parameters to be gamma distributed. Then, for a reasonably high discount factor ( $\delta = 0.95$ ) we have obtained a very good fit, a fact that is indicated by the goodness-of-fit statistics we calculate from the forecast distribution. This seems to indicate that the density approximation we have obtained has an excellent performance when applied to our specific problem.

A possible extension of what is developed here is to consider the same model, for other distributions. Recalling the theory exposed in Section 2.7, we note that the

natural parameter  $\eta_t$  of the Poisson distribution with mean  $\varphi_t$  is given by  $\eta_t = \log \varphi_t$ . Our model links this parameter to the evolution parameter  $\lambda_t$ , for which we assume a normal distribution. As we assume a gamma distribution for  $\varphi_t$ , in order to have conjugacy, we consider, then, a gamma to log-normal approximation. As an extension of this idea, we can think, for example, of binomial data, where the natural parameter  $\eta_t$  is given by the log-odds, so that  $\eta_t = \log(p_t/1 - p_t)$ , where  $p_t$  is the probability of a success. As we usually assume a beta distribution for  $p_t$ , in order to have conjugacy, then, we could develop, for example, an approximation of a beta distribution by a distribution which is called the *logistic normal*, that is, the distribution of  $e^x/(e^x + 1)$ , where  $x$  is normal (Aitchison and Shen, 1980). An idea which is left for future research is to extend this work in general, approximating the conjugate distribution by the appropriate transformation of the normal distribution.

For the model we have developed in Chapter 3, it must be clear that a fundamental role will be played by the sampling interval. The efficiency of the model depends crucially on how frequently we collect information about bookings. For example, daily behaviour of data must present some seasonal pattern (it may include, for example, a trading day effect) that is not presented in the aggregated weekly data. Also, it may well be that one week can represent a very small period of time to collect information for some specific flights, the integrated effect measured by the sum of ten or twelve consecutive weeks being more useful. Therefore, it may be possible that at a larger sampling interval we can obtain a more simple model, since we eliminate the high frequencies effects that will certainly be present when we consider a relatively small period for each observation.

#### 6.4 Review of Chapter 5

With the above ideas in mind, we turn in Chapter 5 to the study of aggregated data. In that chapter, we first analyse the effect of aggregation for the more simple DLM structures (constant first order polynomial and linear growth models) and, then, extend the model developed in Chapter 4 for the case where we have data aggregation.

We begin with the simple constant first order polynomial DLM. For this simple model, we have shown that if conditions (5.4) are satisfied, then, it is still possible for us to represent the aggregated data by the same model. The most important condition is the first one, which involves the relation between the observational and evolution errors, the second condition being linked with the initial information about the parameter. Note that this relation characterizes the presence of high frequency effects in the series. The result of the first theorem of Chapter 5 states that if

we have a large relation between these two variances, then, for a reasonably high aggregation level, we can still have the same simple constant first order polynomial DLM representation for the aggregated data. From (5.7) and (5.8), the new relation will be

$$\frac{V^*}{W^*} = \frac{1}{n^2} \left( \frac{V}{W} - \frac{n^2 - 1}{6} \right) \quad (6.2)$$

and this is smaller than the anterior relation. Then, we expect the discount factor to decrease if the original data can be correctly represented by this simple model. Therefore, we again observe that, if the discount factor increases with aggregation, this indicates that the model we were originally using (for the non-aggregated level) is probably incorrect. The original data, in this case, must present short term irregularities which are subject to long term regularities, the aggregated data providing a better option and allowing for a simpler representation.

It is very important to observe, from (6.2), that the new variance relation can well be zero, depending on  $V$ ,  $W$  and  $n$ . For example, if we have a model where  $W = 2V$ , and we aggregate each  $n = 2$  observations, we can readily see, from (6.2), that we will have  $V^* = 0$ . This result looks surprising, since, in the DLM formulation, we assume the observational and evolution errors to be completely independent. Therefore, we would never expect, in principle, that these errors can be combined in a way that the observational variance will be zero for the aggregated data.

When the conditions (5.4) are not satisfied, the problem can be solved by the introduction of a zero eigenvalue in the system equation. Equivalently, we introduce a non-zero correlation between the observational and evolution errors. This extended model allows us to always have a representation for the aggregated data. We must observe that, for a very high aggregation level, the representation is only possible if we introduce a high (negative) correlation between these errors. However, if we think in terms of forecasting efficiency, we do not lose too much if we use, instead, the random walk model, in the sense that the variance of the one step ahead forecasting error will be only seven per cent higher (the true factor being  $4(2 - \sqrt{3})$ ) in the limit case. This can be easily interpreted, if we think that, after a certain threshold, data are too much gathered together, and the information provided by the last observation will be sufficient to forecast the next one.

Similar results are derived in this chapter when the system equation has the form  $\theta_t = \lambda\theta_{t-1} + \omega_t$ . For this case, (5.89) gives us the limit after which we need to introduce a negative correlation between the errors. It is interesting to see that this limit decreases with a smaller value of  $|\lambda|$ . In other words, if  $|\lambda|$  is smaller, we expect to have a smaller aggregation level  $n$ , after which our model needs sophistication. In

fact, we will not be allowed to aggregate at all, and have the same simple model, if  $|\lambda|$  is very small. That can be easily understood, since a small value of  $|\lambda|$  represents a fast loss of information on the parameter, and, therefore, we will have dominance of the random term in the system equation. Hence, the simple model will not be sufficient to explain the aggregated data, because much information will be lost in the aggregation process. Then, we have to look for a more sophisticated model, in this case. Again, the introduction of the zero eigenvalue will provide us with a solution. We also have to introduce a high correlation between the errors to have the same form of representation in this case. But, it seems that, also for this model, we are never worse than the case for which  $\lambda = 1$ . Then, the best choice is to choose again the simple random walk after the restriction is broken.

It is important to observe that, also for  $\lambda \neq 1$ , we can aggregate to produce a zero variance observational error, as in the first model. A very simple example is provided by the model for which both errors have unit variance,  $\lambda = 1/2$  and we aggregate each  $n = 2$  observations. Then, from Theorem 4, the aggregated series  $Z_k$  can be represented by the more simple model

$$Z_k = \frac{1}{4}Z_{k-1} + \delta_k$$

where  $\delta_k$  is a sequence of independent errors with  $Var[\delta_k] = 45/8$ . Hence, we again have a simpler model by aggregating.

For the general TSDLM model, with a  $p \times p$  system matrix  $G$  it will be always possible to find a representation for the aggregated data, by introducing the zero eigenvalue in the system equation. This is simply because we can have, at least, the state space representation given in the beginning of Section 5.5. Hence, we would like to know when this zero eigenvalue does not need to be introduced, or, equivalently, when we can still maintain the same simple representation, with independent errors, for the aggregated data. The answer is that this will be possible, whenever the following two conditions are satisfied (assuming we do not have to worry about the initial information about the state; for example, we can assume that the initial variance is very large): the first condition is that we guarantee that the expression giving the variance of the observational error in our representation must be non-negative, and the second condition states that the covariance matrix of the evolution errors in the system equation must be non-negative definite. This last condition means that all principal minors of the covariance matrix of the evolution errors must be non-negative. These  $p$  principal minors will be, in general, complicated polynomial expressions involving the aggregation level  $n$ , the observational error

variance  $V$  of the original model and the variances and covariances of the errors in the system equations of the original model. We can try, however, for some simple particular cases, obtain conditions for maintainance of the simple model, when such expressions will be relatively simple.

We have, then, studied the linear growth structure with unit eigenvalues, where the system matrix  $G$  is  $J_2(1)$ . For this model, some interesting results can be obtained. First, it is interesting to observe, from equations (5.185) to (5.187) that the quantities of note are  $V$ ,  $W_1 - W_{12}$ ,  $W_{12}$  and  $W_2$ . Observe that (5.186) gives us  $W_1^* - W_{12}^*$ . It is important to analyse the case where the errors in the system equation are originally uncorrelated. For this case, we have concluded that, when  $n = 2$ , we just have to worry about one condition and that this condition corresponds exactly to the condition we had in the simple first order model,  $2V \geq W$ . Then, there is no fundamental difference between the linear growth structure and the more simple first order model, when the errors are uncorrelated, in the sense that exactly the same condition must be satisfied for aggregation. An interesting subject of study here is to search for a possible generalization of this result, and we ask if a similar phenomenon occurs when we aggregate each  $n = 3$  observations in the quadratic model with uncorrelated errors.

A very important observation is that, still here, the model can be drastically reduced when we aggregate. As a simple example, we can see, from (5.186), that we can have a representation with uncorrelated errors in the system equation, if, in the original model,  $W_1$  is very large. Another interesting numerical example is that one at the end of Section 5.6. In that example we have verified that the complex linear growth structure of the original model can be reduced to the very simple random walk structure by data aggregation. This again is a clear indication that the good choice of the sampling interval works as a preponderant factor to arrive at a good model which is consistent with the main interests of the forecaster.

The last topic that is treated is the application of the ideas introduced in this chapter to the booking problem we have studied in Chapter 4. Our basic motivation is the idea that one week can be a relatively small period of time for us to collect information about the booking of some particular flights. A number of different factors can be responsible for the weekly variation of the bookings. However, it may be reasonable to suppose, based on the observed regularity of past data, that these many different factors, which influence the bookings for each week, like reduction in prices, weather conditions, and others, combine to produce a simple effect over a greater sampling period. The consequence this brings for the model is that we need

a reduced set of parameters when we work with this greater sampling period, and we can more easily identify the long term behaviour of the process, which must be our first aim. This long term behaviour may be hidden by the high frequency effects the are present in the original model, effects that are cancelled when we turn to the aggregated observations.

We can develop our model by considering two types of aggregation. The first one is aggregation along each booking period. This case is easily dealt with. Since we work with independent Poisson distributions, the aggregated observations are still independent Poisson random variables. This sort of aggregation can contribute to filter high variation of data, especially if we work with first class bookings, where the number of persons booking each week is very small. A second type of aggregation corresponds to aggregate data for different flights. This is a much more difficult approach, and we have to consider the evolution of the parameter which gives the demand level for each flight. This type of aggregation may be important, though, since it contributes to eliminate the cross flight effects, and it is left for future reasearch.

## BIBLIOGRAPHY

- Abramowitz, M., and Stegun, I.A., "Handbook of Mathematical Functions," Dover, New York, 1965.
- Aitchison, J., and Dunsmore, I.R., "Statistical Prediction Analysis," Cambridge University Press, Cambridge, 1975.
- Aitchison, J., and Ho, C.H., *The multivariate Poisson-log normal distribution*, Biometrika **76**, 4 (1989), 643-53.
- Aitchison, J., and Shen, S.M., *Logistic-normal distributions: Some properties and uses*, Biometrika **67**, 2 (1980), 261-72.
- Akaike, H., *Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes*, Annals of the Institute of Statistical Mathematics **26** (1974), 363-387.
- Ameen, J.R.M., and Harrison, P.J., *Normal discount Bayesian models*, Bayesian Statistics **2** (1985), 271-298.
- Amemiya, T., and Wu, R.Y., *The effect of aggregation on prediction in the autoregressive model*, J. Amer. Statist. Ass. **339** (1972), 628-632.
- Anderson, B.D.O., and Moore, J.B., "Optimal Filtering," Prentice-Hall, New Jersey, 1979.
- Astrom, K.J., "Introduction to Stochastic Control Theory," Academic Press, New York, 1970.
- Azzalini, A., *Approximate filtering of parameter driven processes*, J. Time Series Analysis **3** (1983), 219-224.
- Bernardo, J.M., *Reference posterior distributions for Bayesian inference (with discussion)*, J. R. Statist. Soc. (Ser. B) **41** (1979), 113-148.
- Box, G.E.P., and Cox, D.R., *An analysis of transformations (with discussion)*, J. R. Statist. Soc. (Ser. B) **26** (1964), 241-252.
- Box, G.E.P., and Draper, N.R., "Evolutionary Operation," Wiley, New York, 1969.
- Box, G.E.P., and Jenkins, G.M., "Time Series Analysis: Forecasting and Control," Holden-Day, San Francisco, 1976.
- Box, G.E.P., and Tiao, G.C., "Bayesian Inference in Statistical Analysis," Addison-Wesley, Massachusetts, 1973.
- Brewer, K.W., *Some consequences of temporal aggregation and systematic sampling for ARMA and ARMAX models*, Journal of Econometrics **1** (1973), 133-154.
- Brown, R.G., "Statistical Forecasting for Inventory Control," McGraw-Hill, New York, 1959.



- Brown, R.G., "Smoothing, Forecasting and Prediction of Discrete Time Series," Prentice Hall, New Jersey, 1963.
- Bulmer, M.G., *On fitting the Poisson-lognormal distribution to species abundance data*, Biometrics **30** (1974), 101-10.
- Conte, S.D., and de Boor, C., "Elementary Numerical Analysis (An Algorithmic Approach)," McGraw-Hill, 1981.
- Crow, E., and Shimizu, K., "Lognormal Distributions, Theory and Applications," M. Dekker, New York, 1988.
- DeGroot, M. H., "Probability and Statistics," Addison-Wesley Publishing Company, Massachusetts, 1975.
- Duncan, D.B., and Horne, S.D., *Linear dynamic regression from the viewpoint of regression analysis*, J. Amer. Statist. Ass. **67** (1972), 815-821.
- Fuller, W.A., "Introduction to Statistical Time Series," John Wiley & Sons, New York, 1976.
- Godolphin, E.J., and Harrison, P.J., *Equivalence theorems for polynomial-projecting predictors*, J. R. Statist. Soc. (Ser. B) **37** (1975), 205-215.
- Godolphin, E.J., and Stone, J.M., *On the structural representation for polynomial-projecting predictor models based on the Kalman filter*, J. R. Statist. Soc. (Ser. B) **42** (1980), 35-45.
- Goldstein, M., *Bayesian analysis of regression problems*, Biometrika **63** (1976), 51-58.
- Gonzalez, P., *Temporal aggregation and systematic sampling in structural time-series models*, Journal of Forecasting **11** (1992), 271-281.
- Graybill, F., "Matrices with Applications in Statistics," Wadsworth, California, 1983.
- Green, M. and Harrison, P.J., *On aggregate forecasting*, Research Report 2 (1972); Department of Statistics, University of Warwick.
- Harrison, P.J., *Short-term sales forecasting*, Applied Statistics **15** (1965), 102-139.
- Harrison, P.J., *Exponential smoothing and short-term forecasting*, Man. Sci. **13** (1967), 821-842.
- Harrison, P.J., *Convergence for dynamic linear models*, Research Report 67 (1985); Department of Statistics, University of Warwick.
- Harrison, P.J., *Bayesian forecasting in O.R.*, Operational Research (1988); N.B. Cook and A.M. Johnson (Eds.). Pergamon Press, Oxford.
- Harrison, P.J., and Akram, M., *Generalized exponentially weighted regression and parsimonious dynamic linear modelling*, Time Series Analysis: Theory and

- Practice **3** (1983); (O.D. Anderson, ed.) North-Holland (with discussion).
- Harrison, P.J., and Stevens, C.F., *A Bayesian approach to short-term forecasting*, Oper. Res. Quart. **22** (1971), 341-362.
- Harrison, P.J., and Stevens, C.F., *Bayesian forecasting (with discussion)*, J. R. Statist. Soc. (Ser. B) **38** (1976), 205-247.
- Harrison, P.J., and West, M., *Practical Bayesian forecasting*, The Statistician **36** (19), 115-125.
- Hartigan, J.A., *Linear Bayesian methods*, J. R. Statist. Soc. (Ser. B) **31** (1969), 446-454.
- Harvey, A.C., "Time Series Models," Philip Allan, Oxford, 1981.
- Holt, C.C., *Forecasting seasonals and trends by exponentially weighted moving averages*, O. N. R. Research Memo. **52** (1957); Carnegie Institute of Technology.
- Jazwinski, A.H., "Stochastic Processes and Filtering Theory," Academic Press, New York, 1970.
- Kalman, R.E., *A new approach to linear filtering and prediction problems*, J. of Basic Engineering **82** (1960), 35-45.
- Lee, A.O., *Airline reservations forecasting*, 28th Annual AGIFORS Conference (1988), 171-198.
- Luenberger, D.G., "Introduction to Linear and Nonlinear Programming," Addison-Wesley, 1973.
- Luenberger, D.G., "Introduction to Dynamic Systems: Theory, Models and Applications," John Wiley & Sons, New York, 1979.
- Magnus, J. and Neudecker, H., "Matrix Differential Calculus," John Wiley & Sons, New York, 1988.
- McCullagh, P., and Nelder, J.A., "Generalized Linear Models," Chapman and Hall, London, 1983.
- Migon, H.S., and Harrison, P.J., *An application of non-linear Bayesian forecasting to television advertising*, Bayesian Statistics **2** (1985); J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (Eds.). North Holland, Amsterdam, and Valencia University Press.
- Morrison, N., "Introduction to Sequential Smoothing and Prediction," McGraw Hill, 1969.
- Nelder, J.A., and Wedderburn, R.W.M., *Generalised linear models*, J. R. Statist. Soc. (Ser. A) **135** (1972), 370-384.
- Ord, J.K., "Families of Frequency Distribution," Griffin, London, 1972.
- Pole, A., and West, M., *Reference analysis of the dynamic linear model*, Journal

- of Time Series Analysis **10** (1989), 131-147.
- Preston, F.W., *The commonness, and rarity, of species*, Ecology **29** (1948), 254-83.
- Rao, C.R., *A note on a generalized inverse of a matrix with applications to problems in mathematical statistics*, J. R. Statist. Soc. (Ser. B) **24** (1962), 152-158.
- Rao, C.R., "Linear Statistical Inference and Its Applications," John Wiley & Sons, New York, 1973.
- Rao, S.S., "Optimization, Theory and Applications," Wiley Eastern, New Delhi, 1978.
- Reid, D.D., *The Poisson lognormal distribution and its use as a model of plankton aggregation*, Statistical Distributions in Scientific Work (1981); Ed. C. Taillie, G.P. Patil and B. Baldessari, **6**, 303-16; Dordrecht, Holland: Reidel.
- Rothstein, M., *O.R. and the airline overbooking problem*, Operations Research **33** (1985), 237-248.
- Seber, G.A.F., "Linear Regression Analysis," John Wiley & Sons, New York, 1977.
- Smith, A.F.M., and West, M., *Monitoring renal transplants: An application of the multi-process Kalman filter*, Biometrics **39** (1983), 867-878.
- Smith, J.Q., *A generalization of the Bayesian steady forecasting model*, J. R. Statist. Soc. (Ser. B) **41** (1979), 378-387.
- Smith, J.Q., *A comparison of the characteristics of some Bayesian forecasting models*, International Statistical Review **60** (1992), 75-87.
- Smith, R.L., and Miller, J.E., *Predictive records*, J. R. Statist. Soc. (Ser. B) **48** (1986), 79-88.
- Souza, R.C., *A Bayesian-entropy approach to forecasting: the multi-state model*, Time Series Analysis (1981); O.D. Anderson (Ed.). North-Holland, Houston, Texas.
- Stram, D.O., and Wei, W.W.S., *Temporal aggregation in the ARIMA process*, Journal of Time Series Analysis **7** (1986), 279-292.
- Theil, H., "Principles of Econometrics," Wiley, New York, 1971.
- Tiao, G.C., *Asymptotic behavior of time series aggregates*, Biometrika **59** (1972), 523-531.
- Wei, W.W.S., *Some consequences of temporal aggregation in seasonal time series models*, Seasonal Analysis of Economic Time Series (1978), 433-448; edited by A. Zellner, Government Printing Office: Washington D.C.
- West, M., and Harrison, P.J., *Monitoring and adaptation in Bayesian forecasting models*, J. Amer. Statist. Ass. **81** (1986), 741-750.
- West, M., and Harrison, P.J., "Bayesian Forecasting and Dynamic Models," Springer,

New York, 1989.

West, M., Harrison, P.J., and Migon, H.S., *Dynamic generalised linear models and Bayesian forecasting (with discussion)*, J. Amer. Statist. Ass. **80** (1985), 73-97.

West, M., Harrison, P.J., and Pole, A., *BATS : A user guide*, Research Report 114 (1987); Department of Statistics, University of Warwick.

Young, P.C., "Recursive Estimation and Time Series Analysis," Springer-Verlag, Berlin, 1984.

Zellner, A., "An Introduction to Bayesian Inference in Econometrics," John Wiley and Sons, New York, 1971.